JAIST Repository
https://dspace.jaist.ac.jp/

Title	Simulation-based Data Mining Solution to the Structure of Water Surrounding Proteins
Author(s)	Dam, Hieu Chi; Ho, Tu Bao; Sugiyama, Ayumu
Citation	Proceedings of the Twenty-Second International Join Conference on Artificial Intelligence (IJCAI-2011): 2424-2429
Issue Date	2011
Type	Conference Paper
Text version	publisher
URL	http://hdl.handle.net/10119/10332
Rights	Copyright (C) 2011 International Joint Conferences on Artificial Intelligence. Hieu Chi Dam, Tu Bao Ho, Ayumu Sugiyama, Proceedings of the Twenty-Second International Join Conference on Artificial Intelligence (IJCAI-2011), 2011, 2424-2429. http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-404
Description	

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Simulation-Based Data Mining Solution to the Structure of Water Surrounding Proteins

Hieu Chi Dam[1,3,4], **Tu Bao Ho**[1,2], **Ayumu Sugiyama**[1,4]
[1]School of Knowledge Science, Japan Advanced Institute of Science and Technology, Japan
[2]John von Neumann Institute, Vietnam National University at Ho Chi Minh City, Vietnam
[3]University of Science, Vietnam National University at Ha Noi, Vietnam
[4]ERATO, Japan Science and Technology Agency, Japan
{dam, bao}@jaist.ac.jp, ayumu@ishikawa-sp.com

Abstract

What is structure of water surrounding proteins remains as one of fundamental unsolved problems of science. Methods in biophysics only provide qualitative description of the structure and thus clarifying the collective phenomena of a huge number of water molecules is still beyond intuition in biophysics. We introduce a simulation-based data mining approach that quantitatively model the structure of water surrounding a protein as clusters of water molecules having similar moving behavior. The paper presents and explains how the advances of AI technique can potentially solve this challenging data-intensive problem.

1 Introduction

It is well known that water plays an essential role in biological systems. In short, biological molecules function in the water environment and water influences on the molecules' functions through their interactions. Water properties in biological systems have been studied for well over a century by a wide range of physical techniques, but progress has been slow and erratic [Halle, 2004].

On its 125th anniversary in 2005, the Science magazine had a project of choosing 125 unsolved problems of science based on criteria: how fundamental they are, how broad-ranging, and whether their solutions will impact other scientific disciplines [Kennedy and Norman, 2005]. One of these problems is *what is structure of water*? A special and important case of the above problem is structure of water surrounding proteins (water in protein solutions). It is well known that the protein folding (another unsolved problem of science) and protein functions strongly depend on the structure of water in which they interact with the protein. The hydration water–water molecules at the protein surface those most influence on proteins–has received much attention from researchers, but also has been a rich source of controversy and confusion, e.g., [Israelachvili and Wennerstrm, 1996], [Svergun *et al.*, 1998], [Pizzituttu *et al.*, 2007], [Zhang *et al.*, 2009].

Broadly, water molecules in protein solutions are qualitatively classified into three categories: (i) *bound water* that strongly bound to the protein (often internally in protein cavities and thus also called cavity water), (ii) *hydration water* at the protein surface that has direct interactions with the protein (also called surface water), and (iii) *bulk water* that surrounds the protein at a separation and is not in direct contact with the protein. It is well known that the three water categories have different functions. Individually bound water has multiple contacts that stabilize the protein structure. Hydration water has heterogeneous dynamical behavior, contributing to protein folding, stability and dynamics, and interacting with the bulk water. Bulk water is free to move and continuously exchanges with hydration water, and indirectly influences on the protein [Bizzarri and Cannistraro, 2002], [Halle, 2004].

Much effort has been devoted to quantitatively model the relative motion (orientation, rotation and velocity) and dynamical properties of individual water molecules in protein solutions [Bizzarri and Cannistraro, 2002]. However, methods in biophysics do not allow us to quantitatively describe the structure of collective motion, which is essential, of water molecules surrounding a protein. Until recently, the structure of hydration water is only broadly and statically described by layers with fixed distance from the protein surface to water molecules. For example, the structure is described with two layers in [Chen *et al.*, 2008], the first hydration layer includes water molecules within 2.75Å from the protein surface, and the second one from 2.75Å to 4.5Å, respectively. It is worth noting that such models of water structure do not rationally reflect the nature of heterogenous moving behavior of water molecules near the protein surface.

Motivated by using data mining to solve data-intensive problems in other sciences, we aim to model the structure of water molecules with collective motions at the protein surface. The essence in overcoming the limitation of the currently known static structure is to model the dynamic moving behavior of water molecules that allow us to imply their interactions with other molecules including proteins, and to characterize the different moving behaviors of hydration water and bulk water. The key idea is to define the structure of water surrounding a protein as dynamics clusters each consists of molecules having similar moving behavior. This idea is computable by studying it *in silico* via simulation and data mining. By appropriate molecular dynamics (MD) simulation of water in protein solution we generated a huge data volume of 18.77 terabytes and by analyzing such data we can determine the dynamic structure of water in protein solution.

The main contributions of this work include: (1) a quanti-

tative and rational solution to the structure of water molecules surrounding proteins; (2) a simulation-based data mining approach that can be applied to certain scientific domains where simulation plus mining MD simulated data is the sole way to do the research.

In the rest of this paper, Section 2 summarizes some related work, the problem and our framework, Section 3 presents the modeling of structure of water surrounding proteins, and Section 4 concludes the work.

2 Related work and problem statement

2.1 Related work on water structure

Structure and dynamics of water molecules are intrinsically related and often considered together. Different physical methods to investigate dynamical and structural properties of individual hydration water molecules have been developed as surveyed in [Bizzarri and Cannistraro, 2002].

[Tarek and Tobias, 2000] shown that hydration water molecules are much less structurally defined than bound water molecules, and are much more mobile, with residence times on the order of tens of picoseconds. Also, the rotational motion of hydration water is slowed down by about a factor of five compared to bulk water.

[Smith *et al.*, 2004] found that the average density of the first hydration layer (0-3Å from the surface) is significantly higher than that of bulk water. About two-thirds of the first-layer density increase of *ca.* 15% is due to this geometric contribution. The remaining one-third (5%) density increase involves significant changes in the average water structure.

Recently, [Chen *et al.*, 2008] characterized the first hydration water layer between protein and water by the first SDF (surface distribution function) and RDF (radial distribution function) maxima at a radius of 2.75Å and 2.75Å, respectively, and the second layer of hydration water by the second SDF and RDF maxima at 3.65Å and 4.5Å, respectively.

It was shown that the H-bond water network in the hydration layer is much more rigid than in bulk water and the relaxation in a few picoseconds must be from local collective water-network motions [Zhang *et al.*, 2009]. Motion of water molecules can only be studied *in silico* by simulation, but there is so far no computational methods for modeling the structure of water surrounding proteins.

2.2 Problem and simulation-based data mining

While data mining has been widely used in computational biology, it has also been applied in physics and chemistry. In their early works, [Curtarolo *et al.*, 2003], [Fischer *et al.*, 2006], the authors merged data mining with the first principle calculations based on density functional theory and improved by principle component analysis (PCA) to predict the crystal structure of metal alloy series's. Not only for prediction task, data mining is expected to bring insight to various problems which seem impossible to solve with current popular methods of the fields where reformative ideas are needed.

Recently, there have been some work on combining simulation and data mining for solving scientific and engineering problems, such as market research [Better *et al.*, 2007] or aircraft engine fleet management [Painter *et al.*, 2006], defect structure in DM simulation data [Mehta *et al.*, 2005].

1. Carry out MD simulations of the motion of water molecules in interaction with proteins: the *moving* of each water molecule is measured in a time interval Δt.
2. Transform the moving data into a new feature space that represents *moving behavior* of the water molecules.
3. Develop an appropriate clustering method to model the structure of water in protein solution as clusters by water molecule moving behavior. These clusters in a given time interval Δt are called *static structure*.
4. Investigate the *dynamics* of hydration water and the *dynamic structure* of water surrounding proteins from the stream of static structures created in consecutive time intervals by steps 1-3.

Figure 1: Framework of simulation based-data mining to modeling structure of water surrounding proteins.

Noting that MD simulations of biomolecules have remarkably progressed with extremely high precision by using accurate force field and high computation performance by parallelization techniques that made possible the investigation of their complex motions that are not accessible from experiment. The study of water properties in biological systems is intensively based on simulation. To our best knowledge, there is no work using data mining to investigate high precisely simulated data for macromolecules in life science.

The problem stated in our work is *to quantitatively model structure of water surrounding proteins as well as to characterize the collective motions of water molecules in their interactions with the others including the proteins.*

To this end, we develop a simulation-based data mining approach that exploits the data obtained by high quality MD simulation to create a quantitative structure of water in protein solutions. It is important to note that there is a causal relation between motion of each water molecule and its interaction with other molecules in the simulation, which follows precisely the fundamental rules of physics. However, clarifying the collective phenomena of a huge number of water molecules is still beyond intuition in physics, and thus our approach can be made possible with assumptions based on our domain knowledge.

Based on our assumption that the dynamical behavior of bulk water is different from that of non-bulk water molecules, our idea is to define the water structure as three clusters each contains a mass of water molecules having the similar moving behaviors in a time interval. This idea is computable by studying it *in silico* via simulation and data mining.

Figure 1 presents our framework consisting of four steps. The first three ones consist of the simulation of the moving and interactions of water in protein solutions, the transformation of moving data into a feature space representing water moving behavior and the static structure model of water surrounding proteins are the objective of this work. The fourth step on the dynamic structure will be further investigated.

Table 1: Detail information of three kinds of protein and water

PDB ID	1HEL	4PTI	1PSV	Water
# protein atoms	1968	898	479	0
# protein residues	159	58	28	0
Total ions	8Cl-	6Cl-	3Cl-	0
# solvent water	11432	7993	6104	12071
# total atoms	36272	24883	18798	36213
Water box	(69.31, 68.78, 76.53)	(64.76, 60.30, 64.17)	(60.64, 59.19, 42.98)	(71.55, 1.42, 71.94)

Table 2: Size and volume of the simulation data

Simulation Model	1HEL	4PTI	1PSV	Water
# Reduced data points	30.8×10^6	21.6×10^6	16.5×10^6	32.6×10^6
Original data volume	5.86 TB	4.02 TB	3.04 TB	5.85 TB

3 Structure of water surrounding proteins

3.1 MD simulation and simulated data

We have carried our MD simulations with 3 kinds of proteins, concretely lysozyme (1HEL), trypsin (4PTI) and computationally designed peptide (1PSV) which are well known in the literature. The detail information of these three proteins is given in Table 1.

MD simulations are performed for three protein solutions by AMBER10 program package with force field 03 [Pearlman *et al.*, 1995]. Three simulations are carried out for constructing the water data under protein solutions. The simulations of proteins in the water environment are carried out using the structure of 1HEL, 4PTI and 1PSV.

All of the simulations are designed in constant volume condition with periodic boundary. To avoid the direct protein–protein interaction from the periodic boundary condition, water thickness was taken at least 15 Å from the protein surface to water box wall. Long range interaction is treated using a cutoff of 12 Å. Internal constraints were relaxed by an energy minimization and following MD simulations under position restraints during 50 picoseconds (ps).

In the simulations, each water molecule (contains 2 hydrogen atoms and 1 oxygen atom) in a system at a step of 1 femtosecond (fs) is described by 18 values (6 values for each atom). Each system, for example, for the bulk water total 12071 water molecules will be described by $18 \times 12071 \simeq 2.1 \times 10^5$ values. The variations of all of these values are simulated in an interval of 3.75 nanoseconds (ns), i.e., 3.75×10^6 steps which are consequently grouped into 150 subintervals Δt of length 25 picoseconds (ps), each contains 25000 steps. The total number of values for the whole interval and the number of values for each subinterval we obtain from the simulations, for example, for the bulk water are $18 \times 12071 \times 3.75 \times 10^6 \simeq 8.1 \times 10^{11}$ and $18 \times 12071 \times 25 \times 10^3 \simeq 5.4 \times 10^9$, respectively. Details of the systems, simulation settings, and the obtained data are summarized in Table 2.

3.2 From moving data to moving behavior

As described above, in the obtained raw data, the description of the positions (at each time step) of an water molecule in a subinterval (25 ps) requires $18 \times 25 \times 10^3 \simeq 4.5 \times 10^5$ values and is not appropriate for the modeling purpose. To overcome this difficulty, our key idea is to create a more compact feature space that can represent the moving behavior of water molecules in a time interval, and to map the motion data of each molecule in an interval into one point in this space.

From the simulations, we observe and distinguish several tracks of water molecules via their behavior when moving around as illustrated in Figure 2, and consider three types of moving behavior of water molecules surrounding proteins:

1. Bound behavior: Only moving within a narrow area.
2. Unbound behavior: Moving in a wide area and do not bind anywhere.
3. Changing between unbound behavior and bound behavior during a time interval.

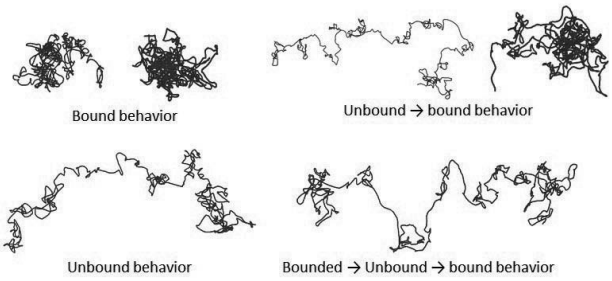

Figure 2: Typical moving behaviors of water molecules in protein solutions.

From the literatures, we learn that the hydrogen bond network relaxation as well as the residence of an water molecule at the surface of a protein occur in a few picoseconds. It is, therefore, natural to consider that the moving behavior, namely, the track of a water molecule in an interval time with an order of ten picoseconds has sufficient information for categorizing that molecule. Our strategy for designing the feature space is to combine the representations of some aspects of how water molecules move.

Moving behavior by a distribution of existence time

We firstly focus on the binding to a specific site of water molecules. The distribution of existence time of an water

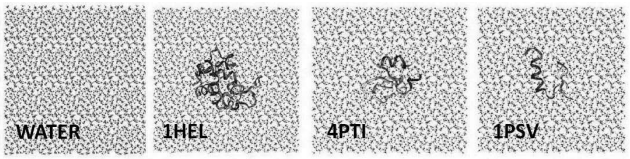

Figure 3: Pure water and water surrounding three proteins. The first left is a W context and the rest are three P contexts used for experiments.

molecule over the moving space in an interval time, with respect to the center of the moving space, will represent this aspect of the moving behavior. We therefore undertake to design features by using the radial distribution of the positions. First, we calculate the distribution from the track of a water molecule i, with parameters r and $\frac{dr}{2}$, in an time interval Δt:

$$G_i^{\Delta t}(r) = P[r - \frac{dr}{2} < |R_i^{t \in \Delta t} - R_{center}^{\Delta t}| < r + \frac{dr}{2}]$$

where, $R_i^{tin\Delta t}$ are the positions at each time step and $R_{center}^{\Delta t}$ is the center (average) position of the molecule in the time interval Δt. Practically, we count the number of MD steps, that the distances from the positions at these steps to the average position of the molecule in the time interval Δt are in between $r - \frac{dr}{2}$ and $r + \frac{dr}{2}$. The distribution is normalized using the total number of time steps in the time interval Δt.

Next, for a simplification, we use a fixed number of points to represent the $G_i^{\Delta t}(r)$ distribution (for various r) of each water molecule, in a time interval Δt. The moving behavior of an water molecule, therefore, can be represented by a point in a multi dimensional vector space $\overrightarrow{G}$:

$$\overrightarrow{G}_i^{\Delta t} = (G_i^{\Delta t}(r_1), ..., G_i^{\Delta t}(r_n))$$

In this study, we choose $n = 25$, and data sets $\overrightarrow{G}$ for the water molecules for all the systems are calculated from the original simulation data.

Moving behavior by a multipole expansion of the track

We secondly focus on the aspect of traveling between sites of water molecules. The idea of multipole expansion can be applied in designing the corresponding features.

We divide the track of a water molecule i in an time interval Δt equally into 2^k parts: the sub-tracks in $\frac{\Delta t}{2^k}$. The $2^k - 1$ distances between centers of every two sub-tracks followed by another are used as features of the moving behavior of the molecule. After carried out the above procedure for $k = 1, ..., m$, we can obtain totaly $\sum_{k=1,...,m}^{m}(2^k - 1) = 2^{m+1} - m - 2$ features ($R_{i,j=1,...,2^{m+1}-m-2}$). The moving behavior of the water molecule i, therefore, can be represented by a point in a multi dimensional vector space $\overrightarrow{R}$:

$$\overrightarrow{R}_i^{\Delta t} = (R_{i,1}^{\Delta t}, ..., R_{i,2^{m+1}-m-2}^{\Delta t})$$

We choose $m = 6$, and data sets $\overrightarrow{R}$ for molecules of all systems are calculated from the original simulation data.

Evaluation of the similarity in moving behavior

Many other representations for the moving behavior of water molecules can be considered. In this paper, we limit ourself to these two representations, and combine them to design the feature space for the next step. The moving behavior of a water molecule i in an time interval Δt, is represented by a point in a multi dimensional vector space $\overrightarrow{F}$:

$$\overrightarrow{F}_i^{\Delta t} = (G_i^{\Delta t}(r_1), ..., G_i^{\Delta t}(r_n), R_{i,1}^{\Delta t}, ..., R_{i,2^{m+1}-m-2}^{\Delta t})$$

The Euclidean distance in vector space $\overrightarrow{F}$ can be used for the similarity evaluation in the moving behavior of water molecules.

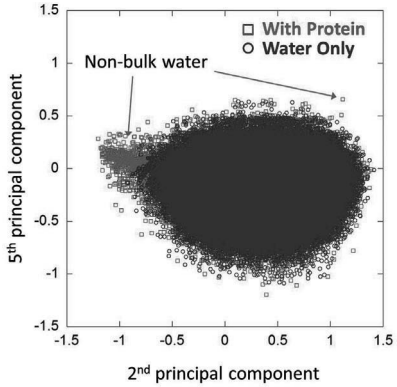

Figure 4: Comparing two contexts of pure water and water in 1PSV protein solution. The non-bulk water molecules can be observed by some principle components.

3.3 Static structure of water surrounding proteins

A *structure* can be viewed as the way in which parts are arranged or put together to form a whole. In this paper we limit ourselves to model the static structure of water surrounding proteins, i.e., the steps 1-3 in the framework (Figure 1).

In the common clustering problem, the objects are usually observed, gathered and clustered from one context (environment). However, in many scientific domains one main interest is to group the objects that have the similar behavior when the context changes. This suggests a new way of clustering by simulating the objects in different contexts in which we can detect their changes. In particular, for the problem of clustering water molecules surrounding a protein, we are interested in the changes of their moving behavior before and after a protein being added into the water. Thus, instead of clustering water molecules on each of three protein contexts (called P contexts) with three proteins 1HEL, 4PTI, 1PSV, we consider one more context of only pure water (called W context) as the first left context in Figure 3 (the data set of the moving behavior of water molecules in this context is calculated in the same way as with the others).

The proposed algorithm is described in Figure 5. key Its idea is to first recognize the non-bulk water molecules from bulk water molecules in a P context then cluster non-bulk water molecules while maintaining bulk water molecules. A

1. Denote the simulated dataset of n_p water molecules from the P context and the simulated dataset of n_w water molecules from the W context, respectively, as:

   $$D^p = \{G_i^p, i = 1, ..., n_p\}, \quad D^w = \{G_i^w, i = 1, ..., n_w\}$$

2. Determine the cluster D_b^p of bulk molecules by checking for each water molecule in C^p if its moving behavior is similar to that of at least one molecule in D^w by using a threshold r_{min}^1. The set $D_{nb}^p = D^p - D_b^p$, consequently, contains non-bulk water molecules of D^p (their moving behavior are called non-bulk behavior).
3. Carry out a hierarchical clustering on D_{nb}^p. The grouping is done when the distance between a cluster to the closest cluster is smaller than a threshold r_{min}^2. The horizontal cutting line is chosen to finally obtain two major clusters with expectation that each contains molecules having similar moving behavior.
4. Examine the variance of coordinates of water molecules in the obtained two major clusters. The cluster with smaller variance is considered as bound cluster and the other as hydration cluster.
5. Count the proportion of the typical moving behavior for each of the bound, hydration and bulk clusters.

Figure 5: Algorithm for detecting clusters of water molecules surrounding proteins.

molecule is non-bulk if its moving behavior is changed when the protein is added to the pure water, i.e., if its moving behavior cannot be observed in the W context. The method is conducted based on the key assumption: *a water molecule in a P context is a bulk molecule if it behaves 'similarly' to at least one water molecule in a W context*. This implies that a water molecule in a P context that do not behave 'similarly' to any water molecule in a W context will be considered as a non-bulk water molecule (steps 1-2).

The non-bulk water molecules are further divided into hydration water molecules and bound water molecules, respectively, by single linkage hierarchical clustering (step 3). The single linkage clustering is chosen as bound water molecules are closer to each other, and the average distance between two closest hydration water molecules is quite larger than that of bound water. Thus, the densities of bound water and hydration water are different. A 10-folds CV was also used to determine the threshold for detecting whether a water molecule in P context behaves similarly to at least one water molecule in W context. Lastly, we model the *static structure* of water surrounding the protein at each simulation interval Δt as *the three groups of clustered molecules each with proportion of their typical type of moving behavior* (steps 4-5). Figure 4 visualizes the non-bulk water distinguished from bulk water by some principal components and Figure 6 illustrates the process of detecting the structure of water in protein solutions.

From a very fundamental principle of physics, we know if two molecules have similar moving behavior, they interact with the environment in a similar way. From biophysics we

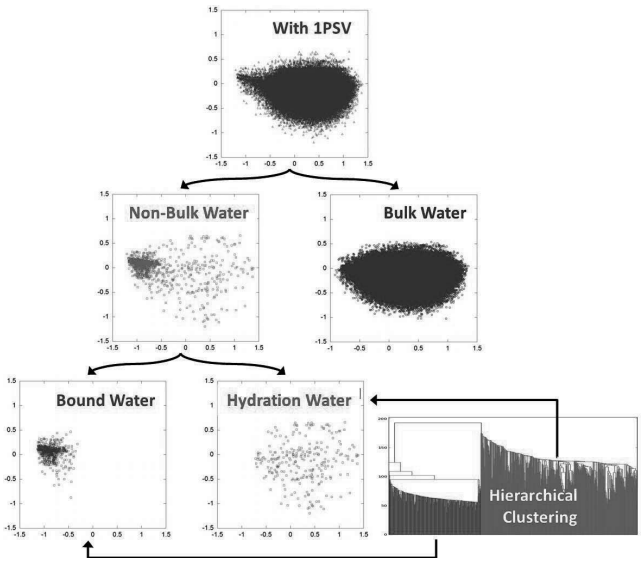

Figure 6: Cluster detection: Water molecules in a protein solution are compared with those in pure water to identify non-bulk and bulk molecules. The non-bulk molecules are then clustered into bound and hydration molecules.

learnt that three categories of water molecules in protein solutions have different functions. Thus, it is natural to believe that the cluster to which a molecule belongs can be determined by (even solely by) how the molecule is moving.

The rationale of structure modeling of water surrounding proteins is evaluated in two criteria. First to compare the radius of moving trace of molecules, in terms of raw simulated data, in each cluster with the 'standard' radius of moving trace (2.3Å, 6.1Å and 9.8Å for the three categories of bound, hydration and bulk, respectively) to count the proportion of correctly clustered molecules. Second to verify physical properties of the water molecules in clusters.

Table 3 shows the results from the evaluation on structures of water molecules surrounding three proteins 1HEL, 4PTI and 1PSV. The columns 2-4 contain the proportions of molecules that are correctly clustered. The proportions in column 2 show that during the time interval in the bound cluster more than 96% of the molecules are only moving in a narrow space with a radius of 2.3Å. Those in column 3 show that in the hydration cluster the change between bound behavior and unbound behavior during the time interval with more than 84% of the molecules are moving in a space with a radius of 6.1Å. The proportions in column 4 show that during the time interval in the bulk cluster 99% of the molecules are moving freely in a space with a radius of 9.8Å.

Though our clustering analysis is based on only the moving behavior of the water molecules, the average numbers in columns 5-8 of non-bulk, bound, and hydration water molecules surround the proteins rationally increase with their size. These demonstrate that the computational results and moving behavior of molecules in the three clusters not only agreed well but also modeled correctly the interactions of a water molecule with other molecules including the protein.

Table 3: Statistics on static structure of water surrounding proteins. Each cell in columns 2–4 contains the proportion of typical type of moving behavior for the corresponding category. The high values of these proportions demonstrate the agreement between the computation results and the domain knowledge in chemical physics about behavior of water surrounding proteins.

Protein	Bound water	Hydration water	Bulk Water	# Residue	# Non-bulk	# Hydration	# Bound
1HEL	0.99	0.84	0.99	159	98.3	56.0	42.3
4PTI	0.98	0.85	0.99	58	39.0	17.6	21.4
1PSV	0.96	0.84	0.99	28	22.2	10.7	11.5

4 Conclusion

By creating a huge volume of simulated data and developing an appropriate clustering method, we proposed a new quantitative description of the water structure in protein solutions. One advantage of the proposed structure is it captures the properties of water known in biophysics and is computable. Furthermore, the simulation-based data mining approach can be applied to a family of problems in certain scientific domains that are similar to the problem considered in this work. Much work should be done to refine and enrich the method, especially to investigate the dynamic structure of water surrounding proteins when considering the stream of static structures over the time.

Acknowledgements

This work is partially supported by Shimoda Nano-Liquid Process Project (ERATO, JST), JSPS project on 'Computational methods for scientific data', and NAFOSTED (Vietnam's National Foundation for Science and Technology Development).

References

[Better *et al.*, 2007] M. Better, F. Glover, and M. Laguna. Advances in analytics: Integrating dynamic data mining with simulation optimization. *IBM Journal of Research and Development*, 51(3/4):1–11, 2007.

[Bizzarri and Cannistraro, 2002] A.R. Bizzarri and S. Cannistraro. Molecular dynamics of water at the protein-solvent interface. *Journal of Physical Chemistry B-Condensed Phase*, 106(26):6617–6633, 2002.

[Chen *et al.*, 2008] X. Chen, I. Weber, and R. Harrison. Hydration water and bulk water in proteins have distinct properties in radial distributions calculated from 105 atomic resolution crystal structures. *Journal Physical Chemistry B*, 112(38):12073–12080, 2008.

[Curtarolo *et al.*, 2003] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder. Predicting crystal structures with data mining of quantum calculations. *Physical review letters*, 91(13):135503, 2003.

[Fischer *et al.*, 2006] C.C. Fischer, K.J. Tibbetts, D. Morgan, and G. Ceder. Predicting crystal structure by merging data mining with quantum mechanics. *Nature Materials*, 5(8):641–646, 2006.

[Halle, 2004] B. Halle. Protein hydration dynamics in solution: a critical survey. *Philo. Transactions of the Royal Society B: Biological Sciences*, 359:1207–1224, 2004.

[Israelachvili and Wennerstrm, 1996] J. Israelachvili and H. Wennerstrm. Role of hydration and water structure in biological and colloidal interactions. *Nature*, 397:219–225, 1996.

[Kennedy and Norman, 2005] D. Kennedy and C. Norman. What don't we know? *Science*, 309(5731):75–102, 2005.

[Mehta *et al.*, 2005] S. Mehta, S. Barr, T. Choy, H. Yang, S. Parthasarathy, R. Machiraju, and J. Wilkins. Dynamic classification of defect structures in molecular dynamics simulation data. *SIAM Data Mining*, pages 161–172, 2005.

[Painter *et al.*, 2006] M.K. Painter, M. Erraguntla, J.L. Hogg, and B. Beachkofski. Using simulation, data mining, and knowledge discovery techniques for optimized aircraft engine fleet management. *Proc. 2006 Winter Simulation Conference*, pages 1253–1260, 2006.

[Pearlman *et al.*, 1995] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman. Amber, a package of computer programs for applying molecular mechanics, normal mode analysis. *Computer Physics Communications*, 91(1-3):1–41, 1995.

[Pizzituttu *et al.*, 2007] F. Pizzituttu, M. Marchi, F. Sterpone, and P.J. Rossky. How protein surfaces induces anormaluos dynamics of hydration water. *Journal Physical Chemistry B*, 111:7584–7590, 2007.

[Smith *et al.*, 2004] J.C. Smith, F. Merzel, A.N. Bondar, A. Tournier, and S. Fischer. Structure, dynamics and reactions of protein hydration water. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359:1181–1190, 2004.

[Svergun *et al.*, 1998] D.I. Svergun, Richard S., M.H. Koch, Z. Sayers, and G. Kuprin, S.and Zaccai. Protein hydration in solution: Experimental observation by x-ray and neutron scattering. *Proceedings of the National Academy of Sciences*, 95:2267–2272, 1998.

[Tarek and Tobias, 2000] M. Tarek and D.J. Tobias. The dynamics of protein hydration water: A quantitative comparison of molecular dynamics simulations and neutron-scattering experiments. *Biophysical*, 79:3244–3257, 2000.

[Zhang *et al.*, 2009] L. Zhang, Y. Yang, Y.T. Kao, L. Wang, and D. Zhong. Protein hydration dynamics and molecular mechanism of coupled water-protein fluatuations. *J. of the American Chemical Society*, 131:10677–106091, 2009.