Published in final edited form as:

IEEE EMBS Int Conf Biomed Health Inform. 2018 March ; 2018: 341–344. doi:10.1109/BHI.

2018.8333438

A General Method for Predicting Amino Acid Residues Experiencing Hydrogen Exchange

Boshen Wang,

Bioinformatics Program, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, USA

Alan Perez-Rathke,

Bioinformatics Program, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, USA

Renhao Li, and

Aflac Cancer and Blood Disorders Center, Department of Pediatrics, Emory University School of Medicine, Atlanta, GA 30322, USA

Jie Liang

Bioinformatics Program, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, USA

Abstract

Information on protein hydrogen exchange can help delineate key regions involved in protein-protein interactions and provides important insight towards determining functional roles of genetic variants and their possible mechanisms in disease processes. Previous studies have shown that the degree of hydrogen exchange is affected by hydrogen bond formations, solvent accessibility, proximity to other residues, and experimental conditions. However, a general predictive method for identifying residues capable of hydrogen exchange transferable to a broad set of proteins is lacking. We have developed a machine learning method based on random forest that can predict whether a residue experiences hydrogen exchange. Using data from the START2FOLD database, which contains information on 13,306 residues (3,790 of which experience hydrogen exchange and 9,516 which do not exchange), our method achieves good performance. Specifically, we achieve an overall out-of-bag (OOB) error, an unbiased estimate of the test set error, of 20.3 percent. Using a randomly selected test data set consisting of 500 residues experiencing hydrogen exchange and 500 which do not, our method achieves an accuracy of 0.79, a recall of 0.74, a precision of 0.82, and an F1 score of 0.78.

I. INTRODUCTION

Proteins are the main molecular players carrying out essential cellular functions such as DNA replication, signal transduction, metabolic catalysis, and transportation of functional molecules. Hydrogen exchange is a widely-used experimental technique for characterizing protein biophysical properties [1]. A class of hydrogen exchange reactions involving hydrogen and deuterium atoms can be measured using mass spectrometry (MS), which can quantify deuterium uptake of protein backbone amide groups. Data gathered from these

hydrogen-deuterium exchange experiments can be used for assessing protein cooperativity, analyzing the dynamics of protein-ligand interactions, and identifying the residues on protein domains involved in binding and protein-protein interactions [2]. For instance, hydrogen exchange mass spectrometry has been applied to explore the mechanism of auto-inhibition of the blood glycoprotein von Willebrand factor (vWF), a protein vital for hemostasis. Results suggest that the short N-terminal region (1238–1260) protects the vWF A1 domain by masking a critical region near the GPIb binding site [3], implying that the vWF N-terminal region may play an important physiological role in hemostatic diseases. Overall, hydrogen exchange information can help delineate key regions involved in protein-protein interactions and to determine functional roles of genetic variants in disease processes.

Previous computational studies on hydrogen exchange have mostly focused on predicting the protection factor, a useful proxy for hydrogen exchange. Vendruscolo *et al* provided a protection factor model for the alpha-lactalbumin protein based on counting the hydrogen bonds formed by backbone amides and the number of proximal residues [4]. Kieseritzky used molecular dynamics (MD) to predict protection factors for cytochrome c [5]. In general, existing computational studies focus on individual proteins and utilize complex, time-consuming methods such as MD simulations. Recently, Vranken utilized a support-vector machine (SVM) algorithm to predict the early-folding residues from the START2FOLD database [6], [7]. They achieved an accuracy of 0.741 on a test data set of 30 proteins, with a precision of 0.361 [6].

In this study, we describe a rapid method for predicting whether a residue experiences hydrogen exchange. Our method is general and can be applied to any protein with known structure. Large scale calculations using our method can provide useful information in analyzing the effects of genetic variants and in assessing potential protein-protein interactions.

II. METHODS

Our method utilizes a random forest classifier to predict if a hydrogen exchange event occurs at a residue of interest. This section provides a summary of our method as well as more detailed descriptions of the features, or model inputs, selected for prediction.

A. Model Design

Figure 1 provides an overview of our method. Our input data sources are the START2FOLD [7] and Protein Data Bank (PDB) [8] databases. START2FOLD provides information on which protein residues experience hydrogen exchange; this is used as the training label as well as for assessing accuracy of our predictions. START2FOLD also provides the environmental conditions under which exchange occurs. The PDB provides three-dimensional atomic positions of the corresponding protein structures. We further compute additional residue-level features such as solvent accessibility and hydrogen bond strength from these positions.

The random forest classifier predicts if a hydrogen exchange event occurs at a residue of interest given its input feature set. Random forest is an ensemble machine learning method

[9]. It consists of a large number of decision trees and the predicted class label is the mode given by this collection of trees. Random forest is robust to overfitting, as each individual tree is trained on only a subset of the data and each decision branch may be limited to only a subset of the features [9]. In our study, we use the “randomForest” R package [10]. For training, we use stratified resampling to present an equal number of positive (exchanged) and negative (non-exchanged) examples to each tree in the forest. As there are many more negative examples compared to positive, stratified resampling helps the forest learn a more sensitive classifier. Our final random forest consists of 1,000 individual trees.

B. Data Collection

From START2FOLD, a comprehensive database of hydrogen exchange experiments [7], we obtain the environmental conditions (pH and temperature) as well as the observed exchangeability of over 13,306 residues from 56 different proteins under 93 experimental conditions. Specifically, 3,790 residues are labeled as exchanged and 9,516 residues are labeled as non-exchanged. For the proteins represented in START2FOLD, we obtain the corresponding three-dimensional atomic positions from the Protein Data Bank.

We have selected features (*i.e.*, inputs) for our hydrogen exchange model based on past experimental evidence. Specifically, these are: solution pH, temperature, backbone amide hydrogen bonding, protein secondary structure, Coulombic interactions, solvent accessibility, and number of proximal residues [11]–[13]. We gather the solution pH and temperature directly from START2FOLD. The remaining features are detailed in the following subsections.

C. Backbone Amide Hydrogen Bonding

Linderstrøm-Lang, a pioneer in hydrogen exchange theory, proposed that only backbone amides are involved in hydrogen exchange events. The standard model proposes that amides exist in either ‘open’ or ‘closed’ states whereby hydrogen exchange can or cannot occur respectively; this model is summarized by the following reaction pathway [14]:

Based on this standard model, we consider only the hydrogen bonds formed between a residue’s backbone amide and any nearby carbonyl oxygens. Since X-ray crystal structures are typically unable to resolve hydrogen atom positions, we use the Reduce [15] program to place these missing atoms. We then define a hydrogen bond to exist if the H-O bond length D is within 3.5 Å and the angle θ formed by the N-H and H-O vectors does not exceed 90 degrees. We define the strength of a single hydrogen bond according to the following formula:

$$\cos \theta / D^2: 0 \leq \theta \leq \pi/2, D \leq 3.5$$

The input features for our model consist of both the number of hydrogen bonds and the summation of hydrogen bond strengths at the residue of interest.

D. Protein Secondary Structure

To determine if a residue experiences hydrogen exchange, we also incorporate information on the secondary structure associated with that residue. Protein secondary structure such as α -helices and β -sheets have characteristic patterns of hydrogen bonds among the backbone atoms. We use the DSSP method [16] to assign a residue to one of the following secondary structure categories: α -helix, 3_{10} -helix, π -helix, β -sheet, hydrogen-bonded turn, or loop; this secondary structure label serves as an input feature to our random forest model.

E. Coulombic Interactions

We use the PROPKA3 [17] method for computing Coulombic interaction strengths at each residue of interest. This feature is only computed when the target residue is one of ASP, GLU, CYS, TYR, HIS, ARG, LYS, C-terminal residue, or N-terminal residue. All other residues are assigned a Coulombic interaction strength of zero.

F. Solvent Accessibility

Solvent accessible (SA) surface area quantifies the amount of atomic surface exposed to solvent. This is an important property as hydrogen-deuterium exchange events occur when the backbone amide is exposed to deuterated solvent [14]. We use the CASTp [18] method for calculating the SA of both the amide nitrogen and the entire residue. In addition, we calculate the normalized residue exposure fraction by dividing the residue SA by the maximum allowed SA as reported by Tien [19]. Lastly, we use DSSP's estimate of the number of water molecules in contact with the residue [16]. These four features are included as inputs to our random forest model.

G. Proximal Residues

Vendruscolo *et al*'s study indicates that the number of proximal residues, referred to as 'protein contacts', may be important for hydrogen exchange [4]. Motivated by this finding, we compute two features related to the number of nearby atoms to the alpha carbon of the residue of interest. The first feature computes the number of alpha carbons within 8 Å, the second feature computes the number of heavy atoms within 15 Å. We compute a third feature based on PROPKA3's buried ratio [17] according to the following:

$$\text{Buried Ratio} = \begin{cases} 0, & \text{for } 0 \leq N < 280 \\ (560 - N)/280, & \text{for } 280 \leq N \leq 560 \\ 1, & \text{for } 560 < N \end{cases}$$

where N represents the total number of heavy atoms within 15 Å of the alpha carbon.

III. RESULTS AND DISCUSSION

A. Performance

Our random forest model has an average out-of-bag (OOB) error of 0.203, with a 0.182 OOB error for non-exchanged residues and a 0.268 OOB error for exchanged residues. Prior to training, we randomly extracted 500 exchanged and 500 non-exchanged residues to serve as the test set. For this test set, our model correctly predicts 371 out of 500 exchanged residues (0.258 error) and 417 out of 500 non-exchanged residues (0.166 error). Our results on this test set are summarized in Table I.

The variable importance plots of Figure 2 describe how each feature affects the performance of the random forest model. Specifically, the mean decrease in accuracy is computed by measuring how exclusion or permutation of a single variable affects the accuracy of the model; permuted variables with larger decreases in accuracy are considered more important. The mean decrease in Gini measures how each variable contributes to the homogeneity of each node in the forest [20]. From Figure 2, we observe that residue type, hydrogen bonding strength, and pH are the most important features utilized by the random forest for predicting exchanged vs non-exchanged residues. Conversely, Coulomb interaction strength and number of hydrogen bonds are considered less important for prediction by the model.

The ROC curves of Figure 3 compare our random forest model to Vranken's SVM early folding model [6]. We created a data set with 326 early folding residues and 326 non-exchanged residues, all from the same set of proteins. The trained Vranken SVM model achieves an area under the curve (AUC) of 0.823. In comparison, our model achieves a marginally higher mean AUC of 0.854 under 5-fold cross validation. We should note that our model utilizes features computed from protein tertiary structure (e.g. hydrogen bonding, electrostatic interactions), which would not be available to Vranken's SVM model as their objective is to infer folding properties.

B. Discussion

Our predictive model is general and achieves a respectable accuracy on residues within the START2FOLD database. Nevertheless, there may be several factors affecting the performance of our model.

1) Experimental Artifacts—Hydrogen exchange events can be captured through either mass spectrometry (MS) or nuclear magnetic resonance (NMR). These different experimental paradigms may introduce specific forms of bias related to parameters such as choice of denaturant and protein concentration. An additional form of experimental bias arises from differences in observation time. Certain proteins in our data set were monitored for relatively short durations. It is possible that if these proteins were monitored for longer durations, more exchange events would be observed; hence, these proteins may be a source of false negatives. In future work, we will extend our model to incorporate the experimental technique and variable observation times (as sufficient data becomes available) in order to control for these possible sources of bias.

2) Hydrogen Bonding—From the variable importance plots of Figure 2, the strength of hydrogen bonding plays a crucial role for predicting hydrogen exchange events. NMR ensembles usually contain explicit hydrogen positions and therefore we can easily compute the hydrogen bonding strength. However, for X-ray crystal structures, the hydrogen positions are unresolved and one must resort to heuristic methods for placing these atoms. Therefore, for X-ray structures, the accuracy of our hydrogen bonding calculations depends on the reliability of the hydrogen addition algorithm. Further, the threshold distance at which hydrogen bonds can occur is a longstanding debate in structural bioinformatics; we use a fairly wide threshold of 3.5 Å to include weak hydrogen bonding. We further assume only the hydrogen bonds formed by the backbone amide contribute to the residue exchange event. It is possible that further optimization of the distance threshold and consideration of additional hydrogen bonds beyond the backbone amide may improve our method.

IV. CONCLUSION

By integrating features based on the structural, energetic, and topological properties of proteins, we have developed a general method that can be applied to any protein with known structure. Our method achieves an accuracy of 0.788 in predicting hydrogen exchange events, at a precision value of 0.817.

Inspection of our model indicates that residue type, pH, and hydrogen bonding strength are the most important features when predicting an exchange event. Compared with molecular dynamics simulation for computing the protection factor, our work shows greater computational efficiency and transferability to a wide range of proteins.

Acknowledgments

This work is supported by NIH grants R01GM079804, R01CA204962, R01GM126558, and R21AI126308.

References

1. Englander SW, Kallenbach NR. Hydrogen exchange and structural dynamics of proteins and nucleic acids. Quarterly reviews of biophysics. 1983; 16(4):521–655. [PubMed: 6204354]
2. Pirrone GF, Jacob RE, Engen JR. Applications of hydrogen/deuterium exchange ms from 2012 to 2014. Analytical chemistry. 2015; 87(1):99. [PubMed: 25398026]
3. Deng W, Wang Y, Lollar P, Li R. A short n-terminal fragment in front of the a1 domain of von willebrand factor prevents the a1 from binding the platelet by masking a region clustered by type 2b vwd mutations. 2016
4. Vendruscolo M, Paci E, Dobson CM, Karplus M. Rare fluctuations of native proteins sampled by equilibrium hydrogen exchange. Journal of the American Chemical Society. 2003; 125(51):15686–15687. [PubMed: 14677926]
5. Kieseritzky G, Morra G, Knapp E-W. Stability and fluctuations of amide hydrogen bonds in a bacterial cytochrome c: a molecular dynamics study. JBIC Journal of Biological Inorganic Chemistry. 2006; 11(1):26–40. [PubMed: 16292670]
6. Raimondi D, Orlando G, Pancsa R, Khan T, Vranken WF. Exploring the sequence-based prediction of folding initiation sites in proteins. Scientific Reports. 2017; 7
7. Pancsa R, Varadi M, Tompa P, Vranken WF. Start2fold: a database of hydrogen/deuterium exchange data on protein folding and stability. Nucleic acids research. 2016; 44(D1):D429–D434. [PubMed: 26582925]

8. Berman HM, Battistuz T, Bhat T, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, et al. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*. 2002; 58(6):899–907. [PubMed: 12037327]
9. Breiman L. Random forests. *Machine learning*. 2001; 45(1):5–32.
10. Liaw A, Wiener M, et al. Classification and regression by random-forest. *R news*. 2002; 2(3):18–22.
11. Preisler R, Mandal C, Englander S, Kallenbach N, Frazier J, Miles H, Howard F. Premelting and the hydrogen-exchange open state in synthetic rna duplexes. *Biopolymers*. 1984; 23(11):2099–2125. [PubMed: 6498293]
12. Balasubramaniam D, Komives EA. Hydrogen-exchange mass spectrometry for the study of intrinsic disorder in proteins. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*. 2013; 1834(6):1202–1209. [PubMed: 23099262]
13. Mandell JG, Falick AM, Komives EA. Measurement of amide hydrogen exchange by maldi-tof mass spectrometry. *Analytical Chemistry*. 1998; 70(19):3987–3995. [PubMed: 9784743]
14. Englander S, Mayne L, Bai Y, Sosnick T. Hydrogen exchange: The modern legacy of linderström-lang. *Protein Science*. 1997; 6(5):1101–1109. [PubMed: 9144782]
15. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of molecular biology*. 1999; 285(4):1735–1747. [PubMed: 9917408]
16. Kabsch W, Sander C. Dssp: definition of secondary structure of proteins given a set of 3d coordinates. *Biopolymers*. 1983; 22:2577–2637. [PubMed: 6667333]
17. Olsson MH, Søndergaard CR, Rostkowski M, Jensen JH. Propka3: consistent treatment of internal and surface residues in empirical p k a predictions. *Journal of chemical theory and computation*. 2011; 7(2):525–537. [PubMed: 26596171]
18. Binkowski TA, Naghibzadeh S, Liang J. Castp: computed atlas of surface topography of proteins. *Nucleic acids research*. 2003; 31(13):3352–3355. [PubMed: 12824325]
19. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum allowed solvent accessibilities of residues in proteins. *PloS one*. 2013; 8(11):e80635. [PubMed: 24278298]
20. Calle ML, Urrea V. Letter to the editor: stability of random forest importance measures. *Briefings in bioinformatics*. 2010; 12(1):86–89. [PubMed: 20360022]

Fig. 1.
Prediction of hydrogen exchanged residues using random forest.

Fig. 2.
Random forest variable importance plots.

Fig. 3.
Receiver operating characteristic (ROC) curve.

TABLE I

Accuracy	0.788
Recall	0.742
Precision	0.817
Specificity	0.834
F1 Score	0.778
Matthews Correlation coefficient	0.578

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript