Development of a tool for the prediction of protein diffusion coefficient using a QSPR model

Supervision: Tiago Picanço Castanheira da Silva, Tim Neijenhuis, Marcel Ottens

Introduction

Protein diffusion coefficients (*D*) are important parameters for the design and analysis of separation and purification processes [1]. This important parameter has been demonstrated to mainly depend on different phenomena, which range from the protein's properties (*e.g.* molecular weight) to environmental conditions (pH, buffer composition, protein concentration, among others) [2]. Several correlations have been proposed to calculate this parameter [3], which are often simplified and do not allow for the accurate estimation of *D*. The determination of protein diffusion coefficients is usually associated with time consuming and/or expensive experiments. A recently developed method at TU Delft [2, 4] uses a H-cell microfluidics device coupled to an in-line UV monitor. This method demonstrated great results, as it allows to use only a fraction of the sample and time that is required for the *D* measurements, compared to other methodologies such as Dynamic Light Scattering (DLS), Taylor Dispersion, Gouy Interferometry, among others.

Alternatively, information derived from the protein amino acid composition, can be used to rationalize the diffusion behavior. Quantitative Structure Property Relationship (QSPR) models aim to relate specific protein features to different phenomena, *e.g.* the diffusion coefficient. Protein features, can be obtained at different complexities. Only using the amino acid sequence, one dimensional (1D) features can be extracted in the form of amino acid composition representing amino acid count, hydrophobicity scores, charge and isoelectric point, among others. Three dimensional (3D) protein models describe the positions of each atom in the system. These can be obtained by experimentation or predicted, *e.g.* using the recently developed AlphaFold2. These models allow to study the protein geometry and its charge and/or hydrophobic distributions. This information is then used to train machine learning (ML) models capable of predicting desired features. Such ML models have already been demonstrated to accurately predict diffusion coefficients for a set of proteins [5]. However, the software used in the aforementioned study is not readily available and therefore difficult to reproduce.

Research Target

During this project we aim to extend this methodology and train a QSPR model to predict the diffusion coefficient using protein features extracted by our in-house software. This python tool represents the protein surface as equally spaced grid points that contain values representing local charge or hydrophobicity (Figure 1). Diffusion coefficients will be determined experimentally and, supplemented with data from literature, will be used to train and validate the QSPR model. This project is composed by two main research targets: i) determination of protein diffusion coefficients with the in-house microfluidics chip of different proteins under different conditions, ii) train and validate the QSPR model to predict the diffusion coefficient of different proteins.

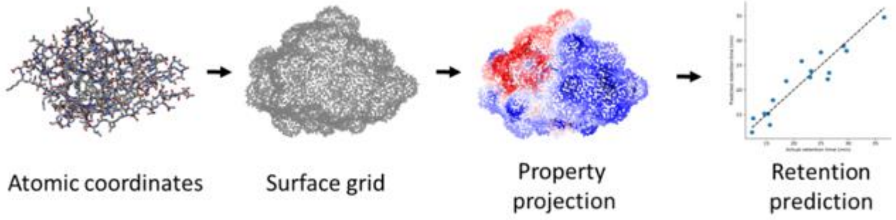

Figure 1 - QSPR workflow

Research Plan

Stage 1: Reviewing relevant literature and learning theoretical background.
Stage 2: Gather relevant data from literature for comparison with the experiments and for training the QSPR model.
Stage 3: Experimental determination of protein diffusion coefficients

3.1 – Replicate protocol previously developed in the group.

3.2 – Study the diffusion coefficients of the one model proteins under different experimental conditions.

3.3 – Determine the diffusion coefficient of the other proteins needed for the QSPR model.

3.4 – Discuss and explain the observations based on diffusion coefficients of the proteins as a function of various factors.

Stage 4: Training and validating a QSPR model that accurately predicts protein diffusion coefficients

4.1 – An analysis of the different calculated protein features, and their correlation to the diffusion coefficient training set.

4.2 – Feature selection and model training using linear regression models.

4.3 – Explore more complex algorithms like support vector machine regression or partial least squares regression.

4.4 – Discuss the features selected to be best for diffusion coefficient prediction for their physical relevant.

Stage 5: Thesis writing and defense preparation

Planning

Task\Month	1	2	3	4	5	6	7	8	9
1. Literature review	🟦								
2. Literature data gathering	🟦	🟦	🟦						
3. Microfluidics experiments	🟦	🟦	🟦	🟦	🟦				
4. QSPR modeling				🟦	🟦	🟦	🟦	🟦	
5. Thesis writing									🟦

References
[1] D. Brune, S.J.P.o.t.N.A.o.S. Kim, Predicting protein diffusion coefficients, 90(9) (1993) 3835-3839.
[2] M. Yu, T.C. Silva, A. van Opstal, S. Romeijn, H.A. Every, W. Jiskoot, G.-J. Witkamp, M. Ottens, The Investigation of Protein Diffusion via H-Cell Microfluidics, Biophysical journal 116(4) (2019) 595-609.
[3] L. He, B.J.B.p. Niemeyer, A novel correlation for protein diffusion coefficients based on molecular weight and radius of gyration, 19(2) (2003) 544-548.
[4] E. Häusler, P. Domagalski, M. Ottens, A.J.C.e.s. Bardow, Microfluidic diffusion measurements: The optimal H-cell, 72 (2012) 45-50.
[5] K.C. Bauer, F. Hämmerling, J. Kittelmann, C. Dürr, F. Görlich, J.J.B. Hubbuch, bioengineering, Influence of structure properties on protein–protein interactions—QSAR modeling of changes in diffusion coefficients, 114(4) (2017) 821-831.