MASARYKOVA UNIVERZITA

Přírodovědecká fakulta

Národní centrum pro výzkum biomolekul

Bakalářská práce

Brno 2021

Stanislav Geidl

MASARYKOVA UNIVERZITA

PŘÍRODOVĚDECKÁ FAKULTA

NÁRODNÍ CENTRUM PRO VÝZKUM BIOMOLEKUL

Title

Bakalářská práce

Stanislav Geidl

Vedoucí práce: prof. RNDr. Jaroslav Koča, DrSc. **Brno 2021**

Bibliografický záznam

Autor:	RNDr. Stanislav Geidl Přírodovědecká fakulta, Masarykova univerzita Národní centrum pro výzkum biomolekul
Název práce:	Title
Studijní program:	Biomolekulární chemie a bioinformatika
Studijní obor:	Studijní obor
Vedoucí práce:	prof. RNDr. Jaroslav Koča, DrSc.
Akademický rok:	2020/2021
Počet stran:	?? + ??
Klíčová slova:	Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo; Klíčové slovo

Bibliographic Entry

Author:	RNDr. Stanislav Geidl Faculty of Science, Masaryk University National Centre for Biomolecular Research
Title of Thesis:	Title
Degree Programme:	Biomolecular chemistry and bioinformatics
Field of Study:	Field of Study
Supervisor:	prof. RNDr. Jaroslav Koča, DrSc.
Academic Year:	2020/2021
Number of Pages:	?? + ??
Keywords:	Keyword; Keyword; Keyword; Keyword; Keyword; Keyword; Keyword; Keyword; Keyword

Abstrakt

V této bakalářské/diplomové/rigorózní práci se věnujeme ...

Abstract

In this thesis we study ...

Místo tohoto listu vložte kopii oficiálního zadání práce bez podpisů.

Poděkování

Na tomto místě bych chtěl(-a) poděkovat ...

Prohlášení

Prohlašuji, že jsem svoji bakalářskou/diplomovou práci vypracoval(-a) samostatně pod vedením vedoucího práce s využitím informačních zdrojů, které jsou v práci citovány.

Prohlašuji, že jsem svoji rigorózní práci vypracoval(-a) samostatně s využitím informačních zdrojů, které jsou v práci citovány.

Brno xx. měsíce 20xx

. . . .

Stanislav Geidl

Contents

Part I

Introduction

Chapter 1

Introduction

In recent years, a vast amount of data about various types of molecules became available. For example, we can obtain the complete human genome of a selected individual in a few days, and about 150 thousand biomacromolecular structures have been determined and published (Protein Data Bank [1]). Furthermore, more than 100 million various small molecules are described in freely accessible databases (e.g., Pubchem [], ZINC [], ChEMBL []). This richness of data caused the formation of novel modern life-science research fields focused on the utilization of this data. The best-known modern life sciences are bioinformatics, structural bioinformatics, systems biology, genomics, proteomics, and also chemoinformatics. These current research specializations have provided many key results in basic and applied research (e.g. [6–12]).

One fascinating and beneficial field utilizing and processing newly available data about small molecules (i.e., drug-like compounds) is chemoinformatics. This discipline offers methodologies for comparing molecular similarity, molecular database search, virtual screening, and the prediction of molecules' properties and activities. This prediction is based on the idea that molecular structures' similarity has a consequence – a similarity in molecular properties. In chemoinformatics, the structure is first described using mathematical characteristics (so-called descriptors) – numbers containing 3D (or 2D or 1D) structure information and applicable as inputs of mathematical models. Then, these models are constructed based on a relation between descriptors and known values of the property or the activity. Such models are called Quantitative Structure-Property Relationship (QSPR) models or Quantitative Structure-Activity Relationship (QSAR) models.

A property, which is strongly required and is therefore often a target of chemoinformatics prediction models is the acid dissociation constant, K_a, and its negative logarithm pK_a. Those pK_a values are of interest in chemical, biological, environmental, and pharmaceutical research [58–60]. pK_a values have found applications in many areas, such as evaluating and optimizing drug candidate molecules, pharmacokinetics, ADME profiling, understanding protein-ligand interactions, etc. Moreover, the critical physicochemical properties such as permeability, lipophilicity, solubility, etc., are pK_a dependent. Unfortunately, experimental pK_a values are available only for a limited set of molecules. In addition to that, obtaining experimental pK_a values for newly designed molecules is very time-consuming because they must be synthesized first. Chemoinformatics approaches for pK_a prediction are therefore currently intensively examined.

For this reason, I also focused on the chemoinformatics way of pK_a prediction in my work. Very promising descriptors for pK_a prediction are partial atomic charges [] because they hold information about the distribution of electron density within the molecule. Specifically, electron densities on atoms close to the dissociating hydrogen provide a clue about its dissociation ability. The most common and accurate method for calculating partial atomic charges is an application of quantum mechanics (QM). QM calculation can be performed via various approaches, introducing different approximation levels (i.e., approximating a wave function by different sets of mathematical equations, which are called basis sets). QM outputs electron distribution in orbitals and this distribution can be divided into individual atoms using several charge calculation schemes (e.g., MPA, NPA, AIM, Hirshfeld, MK, etc.). Therefore, the correlation between pK_a and relevant atomic charges calculated by different QM approaches has been analyzed []. I also focused on this file in my bachelor thesis, developed a workflow for calculation of pK_a using QM partial atomic charges and examined, which types of QM are the most suitable [].

QM charges are accurate, but their calculation is very time-consuming. A faster Alternative to QM charges is empirical charge calculation approaches. Furthermore, if we would like to apply chemoinformatics pK_a prediction models practically – for example, in pre-screening large sets of drug candidates – we need a fast approach. Therefore, in my master thesis, I developed a pK_a prediction workflow based on charges (including Electronegativity Equalization Method).

However, several pieces of the puzzle were still missing. For example, the developed pK_a prediction workflows [] were strongly dependent on 3D structure source, and also, the quality of available EEM charges was low.

Therefore, my dissertation's goal was to develop a workflow that predicts pK_a for molecules not synthesized yet and without available experimental 3D structures.

Specifically, the thesis examined how to improve the process of pK_a prediction via providing suitable inputs. First, the influence of 3D structure source on pK_a prediction accuracy was analyzed. Afterward, the work focused on obtaining high-quality partial atomic charges, which served as descriptors for pK_a calculation. In the end, the authors also support the development of methodology and software tools for obtaining these high-quality charges.

The thesis structure is the following: First, an overview of key fields is provided (Chapter 2), i.e. – 3D structure and approaches for its prediction, charge calculation methods, and pK_a prediction approaches. Next, the achieved results, which we published in three research papers, are briefly described (Chapter 3), and full-texts of the respective published papers are attached in Part ??. During the elaboration of this thesis, I was also involved in other projects. Most of them were not related to pK_a prediction but tightly connected to the field of chemoinformatics or structural bioinformatics. The outcome of these projects consists of several papers and a book I have co-authored. Their title pages are attached in Part ??.

Seznam použité literatury

[1] Berman, H. M., Kleywegt, G. J., Nakamura, H., and Markley, J. L. (2014) The protein data bank archive as an open data resource. *Journal of computer-aided molecular design*, **28**, 1009–1014.