Journal of Biotechnology 261 (2017) 207–214

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

From cheminformatics to structure-based design: Web services and desktop applications based on the NAOMI library

Stefan Bietz[a], Therese Inhester[a], Florian Lauck[a], Kai Sommer[a], Mathias M. von Behren[a], Rainer Fährrolfes[a], Florian Flachsenberg[a], Agnes Meyder[a], Eva Nittinger[a], Thomas Otto[a], Matthias Hilbig[a], Karen T. Schomburg[a,1], Andrea Volkamer[b], Matthias Rarey[a,*]

[a] Universität Hamburg, ZBH – Center for Bioinformatics, Bundesstrasse 43, 20146 Hamburg, Germany
[b] Institute of Physiology, Charité – Universitätsmedizin Berlin, Virchowweg 6, 10117 Berlin, Germany

ARTICLE INFO

Keywords:
Cheminformatics
Structure-based design
Proteins*Plus*
NAOMI

ABSTRACT

Nowadays, computational approaches are an integral part of life science research. Problems related to interpretation of experimental results, data analysis, or visualization tasks highly benefit from the achievements of the digital era. Simulation methods facilitate predictions of physicochemical properties and can assist in understanding macromolecular phenomena. Here, we will give an overview of the methods developed in our group that aim at supporting researchers from all life science areas. Based on state-of-the-art approaches from structural bioinformatics and cheminformatics, we provide software covering a wide range of research questions. Our all-in-one web service platform Proteins*Plus* (http://proteins.plus) offers solutions for pocket and druggability prediction, hydrogen placement, structure quality assessment, ensemble generation, protein–protein interaction classification, and 2D-interaction visualization. Additionally, we provide a software package that contains tools targeting cheminformatics problems like file format conversion, molecule data set processing, SMARTS editing, fragment space enumeration, and ligand-based virtual screening. Furthermore, it also includes structural bioinformatics solutions for inverse screening, binding site alignment, and searching interaction patterns across structure libraries. The software package is available at http://software.zbh.uni-hamburg.de.

1. Introduction

Many biological and medicinal research questions highly benefit from the insights given by protein structure elucidation. Structural biology plays a key role in understanding, utilization, and manipulation of protein function. Therefore, the steadily increasing amount of publically available structures in the Protein Data Bank (PDB) (Berman et al., 2000) constitutes a highly valuable resource for structure-based research. Structural bioinformatics contributes to this field with manifold powerful approaches. Computational methods are involved in structure elucidation, analysis, and quality assessment (Arzt et al., 2005; Goldsmith-Fischman and Honig, 2003; Kleywegt, 2000). Furthermore, they facilitate structure visualization, comparison, and the prediction of macromolecular properties. Preprocessing tools can be used to find appropriate data, to complete structures by missing elements, or to derive knowledge from the structural data that is needed for subsequent applications. Computational simulations like molecular dynamics, docking, and free-energy approximation support the understanding of physiological effects and aim to reduce the amount of necessary but expensive experimental analyses (Leach, 2001). In a similar manner, approaches from cheminformatics assist research on small molecules in areas like medicinal chemistry or biotechnology in general (Gasteiger and Engel, 2006). They support essential data management tasks like the identification of identical compounds, the storage in chemical databases, or filtering by molecular properties. Further applications are file conversion, pattern recognition in sets of similar molecules, or the enumeration of alternative conformations, tautomers, and protonation states. Cheminformatics also covers applications with a more predictive character. Examples are the generation of novel molecules (de novo design) (Schneider and Fechner, 2005) and the prediction of bioactive molecules (ligand-based virtual screening) (Koeppen et al., 2011). Several research questions in life sciences appear exactly at the interface of these two areas of computational science. Structure-based design is one of the key tools in early-phase drug and agrochemical discovery. Also, the development of novel techniques in biocatalysis benefits from this approach (Schneider et al., 2016). In

* Corresponding author.
E-mail address: rarey@zbh.uni-hamburg.de (M. Rarey).
[1] Present address: OSTHUS GmbH, Eisenbahnweg 9-11, 52068 Aachen, Germany.

http://dx.doi.org/10.1016/j.jbiotec.2017.06.004
Received 4 March 2017; Received in revised form 31 May 2017; Accepted 7 June 2017
Available online 11 June 2017

Table 1
Summary of all presented NAOMI-based web services and stand-alone applications.

Web services	Function	Main reference
DoGSiteScorer	Predicts the location of binding pockets and estimates their druggability.	Volkamer et al. (2012)
EDIA	Assesses the conformity of structural atom positions with the experimental electron density.	Nittinger et al. (2015)
HyPPI	Indicates whether protein–protein interactions are permanent, transient, or due to crystallization artefacts.	
PoseView	Draws 2D interaction diagrams of protein–ligand interactions.	Stierand et al. (2006)
Protoss	Adds hydrogen atoms to a macromolecular structure and optimizes their position with respect to polar interactions.	Bietz et al. (2014)
SIENA	Searches alternative binding site structures within the PDB.	Bietz and Rarey (2016)

Stand-alones	Function	Main reference
ASCONA	Calculates alignments of protein binding site conformations.	Bietz and Rarey (2015)
FSees	Enumerates novel molecules from a molecular fragment library.	Lauck and Rarey (2016)
MONA	Facilitates visualization and interactive processing of large molecular data sets.	Hilbig et al. (2013)
mRAISE	Identifies similar molecules via a ligand-based virtual screening approach.	von Behren et al. (2016)
PELIKAN	Searches user-defined protein–ligand interaction patterns in large structural databases.	Inhester et al. (2017)
SMARTSeditor	Supports an interactive design of SMARTS patterns.	Schomburg et al. (2013)
UNICON	Facilitates automatable coordinate generation, sampling of tautomers, protonation states and conformations as well as conversion of different file formats for small organic compounds.	Sommer et al. (2016)

basic research, fields like chemical genomics and metabolomics show strong relationships to both fields. Based on a unique software infrastructure named NAOMI (Urbaczek et al., 2011, 2013, 2014), we have developed a wide range of methods that target problems related to chemistry and structural biology. The NAOMI software library supports cheminformatics and structural biology alike making it an ideal platform for the analysis of protein–ligand complexes. The software tools that we have built on the basis of these methods assist researchers from all areas of life science. Depending on their concrete application range and the respective user community, we either provide the tools as stand-alone software or in context of our web server Proteins*Plus*. In the following, we will briefly introduce all available applications grouped by their implementation strategy and scientific area. An overview of all web services and stand-alone applications is given in Table 1.

2. Web services

Providing the functionality of computational approaches via web services has various advantages over classical stand-alone approaches. Web services are usually the method of choice for providing access to large amounts of preprocessed data. They are platform independent, circumvent installation issues, and are thus accessible to the vast majority of the scientific community. Furthermore, web services mostly employ reduced, easy-to-use interfaces and therefore achieve higher usability and a more intuitive application behavior. For these reasons, we offer web-based solutions for many structure-related research questions. In order to support quick familiarization with the supplied functionality, all of our services are integrated into a single web platform, called Proteins*Plus* (Fährrolfes et al., 2017), offering a unified interface and a standardized workflow for all featured applications. Proteins*Plus* can either operate on a PDB structure (by providing the PDB ID) or on files uploaded by the user (PDB format for macromolecules, SD format for small molecules). The provided structure is visualized as a three-dimensional model using the NGL viewer (Rose and Hildebrand, 2015) (cf. Fig. 1). Its integration into Proteins*Plus* allows several control options including various depiction styles for both protein and ligands, surface visualization, and screenshot generation. The 3D window is also used to illustrate the results for most of the tools integrated into Proteins*Plus*, e.g. binding pockets, predicted hydrogen positions, or electron density fit. Based on a molecule perception algorithm (Urbaczek et al., 2013), ligand molecules are additionally depicted as structure diagrams and SMILES strings. Textual results are presented as sortable tables and can be downloaded for further processing.

Proteins*Plus* covers solutions for multiple problems like structure preprocessing, analysis, and visualization issues as well as the prediction of macromolecular properties. One of the most essential tasks in the context of structure-based research questions is the assessment of the structure's quality. Due to experimental uncertainties and modeling inaccuracies, there might be less experimental evidence for certain parts of a structural model. As this affects the reliability of subsequent interpretations and calculations, it is of great importance for many structure applications to identify such structural uncertainties. In the case of X-ray crystallography, some of these potential error sources can be detected by comparing the experimental electron density with the derived structural model. Various measures have been developed that quantify differences of experimental and modeled structure representations (Jones and Kjeldgaard, 1997; Tickle, 2012) albeit with slightly different purposes. Our recently developed electron density score for individual atoms (EDIA) (Nittinger et al., 2015) aims at the identification of structural elements insufficiently supported by experimental data. In contrast to other methods, this also includes highly flexible substructures, although different uncertainties are still captured with a single consistent measure. Furthermore, EDIA facilitates an atom-wise quality description which, e.g., can be used for an intuitive graphical representation (as included in Proteins*Plus*) or the exclusion of unreliable atoms from conformation-critical analysis strategies.

Another common problem is that most protein structures do not provide a full and precise model of all atoms. Usually, this is due to certain drawbacks of the approaches applied for structure elucidation. For X-ray crystallography, the major issues are the identification of hydrogen positions and the identification of certain side chain orientations, which are both often complicated by insufficient resolution (Davis et al., 2003, 2008). Additionally, numerous structures also lack detailed information on bond orders of atypical residues and ligand molecules. However, many applications dealing with the assessment of molecular interaction like binding affinity estimation or molecular dynamics simulations rely on a complete and accurate atomistic model of the protein. Our hydrogen prediction approach Protoss (Lippert and Rarey, 2009; Bietz et al., 2014) can be used to complete a given structural model by hydrogen atoms, assign unknown bond types, and correct erroneous side chain orientations. In order to achieve an optimal orientation of hydrogen atoms, Protoss optimizes the orientation of rotatable hydrogen atoms and considers alternative protonation states in both ligand and protein moieties. In contrast to this, most competing tools (Brünger and Karplus, 1988; Bass et al., 1992; McDonald and Thornton, 1994, 1995; Hooft et al., 1996; Word et al., 1999; Li et al., 2007; Bayden et al., 2009; Labute, 2009; Krieger et al., 2012) mainly handle those functional groups existing in proteins while neglecting the majority of groups occurring in ligand molecules. This