ДАТА-ОРИЕНТИРОВАННЫЕ ПОДХОДЫ НА ПРИМЕРЕ МОЛЕКУЛЯРНЫХ ДЕСКРИПТОРОВ

DATA-DRIVEN APPROACHES APPLIED ON EXAMPLE OF MOLECULAR DESCRIPTORS

УДК 541.69

DOI: 10.24411/2658-4964-2021-10356

Евлоева Ася Исраиловна, Магистр по Хемоинформатике, Университет ИТМО Россия, Санкт-Петербург

Asya Evloeva, Master degree in Chemoinformatics, ITMO University, Russia, Saint-Petersburg

Аннотация: Статья посвящена анализу важности выбора правильного подхода к кодированию данных. На примере молекулярных дескрипторов рассматривается как данный объект, может быть, представлен в машиночитаемом формате. Оцениваются варианты начиная с квантового уровня молекулярной структуры и заканчивая более общим уровнем, например, молекулярными поверхностями**.**

Annotation: The article is dedicated to review the importance of choosing a proper approach to encode data. On example of molecular descriptors, it analyses how a given object can be presented in a machine-readable format. Observing possibilities starting from quantum level of a molecular structure and ending up with more general level like molecular surfaces.

Ключевые слова: хемоинформатика, молекулярные дескрипторы, машинное обучение, наука о данных, анализ данных.

Key words: chemoinformatics, molecular descriptors, machine learnig, data science, data analysis.

Data science, as an inter-disciplinary field, manages variety of data types over all spheres whether it is sales, marketing, astronomy, and chemistry. In marketing, for instance, one can use sales forecasting or predict a customer churn ratio; in astronomy, it can be pulsar candidate recognition or in the pharmaceuticals, it can be used for drug discovery. There is no significant difference for data science approaches where one takes data from unless this data is presented in a meaningful form for further processing. No secret such diverse data has its own business or science specifics, however it can still be represented in a computer readable way in case to be used with desired machine learning models. The first necessary step is to understand the object of interest (and its properties and particularities) that is going to be used in work. To convert the data into valuable information using machine readable format, one should keep in mind a structure of considered data. A good practice would be using a detailed specification of an object to work with, to consider encoding a variety of possible features of the object, then to select only necessary information for the given task, to approximate in case where it is not sensitive to omit redundant or unnecessary information. In this article, reviewing goes through techniques to encode an object of interest using a machine-readable format on the example of molecules.

Existence of various tasks in chemistry, based on knowledge about molecules, intends to represent molecules in various ways to reach a gain point. Listed here are general approaches to represent molecules for both descriptive and predictive tasks. Relative to that, same logic is applicable to represent sales transaction, pulsars density, and any other object of interest.

Majority of objects that could be represented are most likely to have repeated patterns in data structure/object behaviour. That is what defines a crucial feature to use in further analysis. Music may have repetitions of its chords and limited number of musical notes, pictures have limited colour scheme and may have repetitions of pixels

and groups of pixels, sale transactions may have repetitions of same amount of spends and revenue for some goods, genes may have sequence motifs and so on.

In the case of chemical compounds, there is a set of atoms and a list of their known and possible properties, activities and loads ways of bonding and combining all of them. A description of an object defines its limits. First step is to understand that limitation of data whether it is a musical notation, alphabet letters, chess pieces move, number of colors in color mode, list of possible transactions, or any variety of given object properties. These limits might be approximate or exact, so the idea is to estimate how exact these limits should be for the given task and how deep to go into specification of an object.

For instance, starting from specification of a molecule. Any name of a term like “molecule” is itself meaning of some underlying encoded data. So, what can be encoded by “molecule” in machine readable format and how can it be obtained?

It is known that a molecule consists of atoms and bonds. “A molecule is a particle made up of two or more atoms that are chemically bonded together” [1]. “Atom” is one more term – one more level to describe an object in details. Moreover, there are bonds connecting atoms, “bonds” is a more complex way to describe a molecule.

Here it is a challenge to estimate how deep to go into this specification – is it enough to know just molecule’s atoms to solve a task or is it needed to be calculated on a more detailed level?

Encoding molecules has the same underlying meaning as the idea to «transform data into information and information into knowledge» [Brown, 1998].

To start transforming the data, one should first try to answer some basic questions. How the object of interest can be represented in discrete mathematics? What patterns can be observed and what might be the best mathematical model to describe it? Chosen mathematical models can represent some information of an object, but not always a whole object complexity, so there are usually different approaches to encode the same data.

Therefore, encoding of a molecule can be divided into several levels:

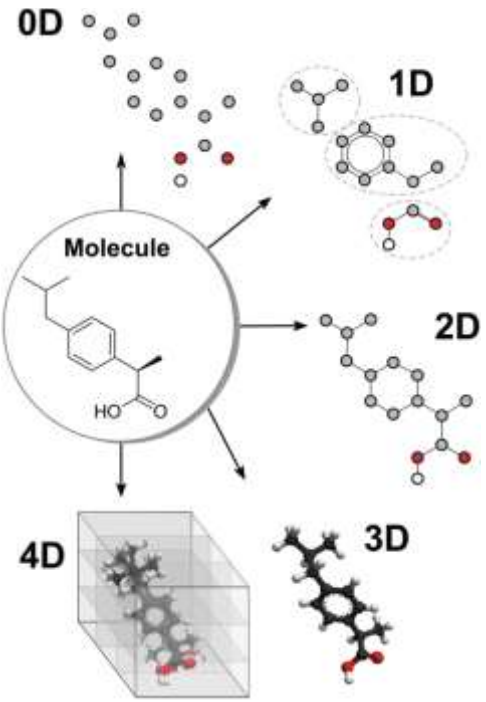


Fig 1. Molecular Descriptors [2]

0D descriptors

This level unites all basic calculations that can be performed on a molecule, for example, atoms counts, bonds counts, molecular weight, molar refractivity, logP (logarithm of the octanol/water partition coefficient).

1D descriptors

A combination of basic counts written as molecular formula, or any fragment counts (e.g. number of OH).

2D descriptors

As is observed, molecular structure in simple words is a system of atoms and their covalent bonds – which strongly reminds a graph model. Using a graph model of a molecule, where atoms are vertices and the covalent bonds are edges it is possible to obtain graph metrics such as distance of a graph (which is called Wiener index in chemoinformatics), total sum of degrees of edges (which is Platt number in chemoinformatics) and any graph representations such as connectivity matrix, adjacency matrix, etc.

3D descriptors

Knowledge that molecule has a complex geometry and can be represented in three-dimensional coordinate system, leads to its 3D spatial properties. To keep initial 3D information of a molecule it is common to use specific chemical file formats, for example, such as SDF (Structure-Data file) which consists of information of atoms and their coordinates.

4D descriptors

4D level of encoding molecules is a combination of atomic coordinates and possible conformations, shape and binding properties. General idea of encoding molecule as a 4D descriptor is that the structure of a molecule reflects its chemical properties which is known as Quantitative Structure–Activity Relationships (QSAR). That principle helps to understand molecular biological activity. An example of 4D descriptors could be molecular surfaces or volumes.

On each level there is a possibility either to use an exact value for each object or to use an approximation. Whether it is atom weight or molecular surface there is always an option to encode given information precisely or approximately. For example, atomic weight is itself a mean of the atomic masses of all isotopes. [3] Second example is that van der Waals radius approach to calculate molecular surface is based on approximation that atoms are spherical even if in fact they are not. Therefore, using a single radius for an atom implies an approximation. [4]

The goal is to estimate when to use approximations and with what level of specification. Knowledge about when it is a crucial point to use or avoid approximations could be a crossroads of whether considered machine learning model would become successful or not.

To sum up, there are loads of molecular descriptors created by now, so the challenge is not to pick any random one of them and to feed a ML model with it. Idea is to keep in mind a final goal and take the best out of given data, meaning – sometimes possibly to develop new descriptor, or to tune its approximation level, as in the end it can lead to more accurate result.

Bibliography

1. Chemicool.com. 2021. Definition of molecule - Chemistry Dictionary. [online] Available at: <https://www.chemicool.com/definition/molecule.html> [Accessed 23 April 2021].
2. Grisoni, F., Consonni, V., & Todeschini, R. (2018). Impact of Molecular Descriptors on Computational Models. Computational Chemogenomics, 171–209.
3. Khan Academy. 2021. Atomic number, atomic mass, and isotopes (article) | Khan Academy. [online] Available at: <https://www.khanacademy.org/science/biology/chemistry--of-life/elements-and-atoms/a/atomic-number-atomic-mass-and-isotopes-article#:~:text=You%20can%20approximate%20the%20atomic,that%20is%20only%20an%20approximation.> [Accessed 23 April 2021].
4. Pubs.acs.org. 2021. Consistent van der Waals Radii for the Whole Main Group. [online] Available at: <https://pubs.acs.org/doi/10.1021/jp8111556> [Accessed 23 April 2021].

Список использованных источников

1. Chemicool.com. 2021. Definition of molecule - Chemistry Dictionary. [Электронный ресурс]: URL: https://www.chemicool.com/definition/molecule.html (дата обращения: 23.04.2021).
2. Grisoni, F.; Consonni, V.; Todeschini, R. Impact of Molecular Descriptors on Computational Models. // Computational Chemogenomics. – 2018. С. 171–209.
3. Khan Academy. 2021. Atomic number, atomic mass, and isotopes (article) | Khan Academy. [Электронный ресурс]: URL: https://www.khanacademy.org/science/biology/chemistry--of-life/elements-and-atoms/a/atomic-number-atomic-mass-and-isotopes-article#:~:text=You%20can%20approximate%20the%20atomic,that%20is%20only%20an%20approximation (дата обращения: 23.04.2021).

4. Pubs.acs.org. 2021. Consistent van der Waals Radii for the Whole Main Group. [Электронный ресурс]: URL: https://pubs.acs.org/doi/10.1021/jp8111556 (дата обращения: 23.04.2021).