STUDY BREAK

Number Crunching towards Molecular Barcoding

Soroush Sardari

Drug Design and Bioinformatics Unit, Medical Biotechnology Dept., Pasteur Institute of Iran, Tehran 13164, Iran

Molecules are the basic and most important players of life. Molecules are composed of atoms, and each molecule has two main specifications or inheritances that are structures, and in biomedical applications, biological function. This dichotomy can be compared to living organisms or humans whom possess genotype and phenotype. Each of the two main inherited elements contains a variety of feature classes and sub-categories. Due to ease of use, the scientific community focuses on elaborating the structure, rather than the function. Of course, this has also caused many debates on the priority of structure over the function or vice versa.

To better understand the features of the structure that could affect function and hence translate them into therapeutic or diagnostic applications, we may look at molecules through their properties that is an intermediary step in the process, which mostly includes the physico-chemical aspects.

Atomic sense of numbers

All molecules are made from atoms. In a simple molecule with two atoms, atoms A and B can bind to form a molecule in two ways AB and BA. Linearly speaking, both molecules are the same, but if there are two fragments, with asymmetry or different binding points, the two molecules would be different.

Now, if we have three atoms A, B, and C, there are options ABC, ACB, BAC, BCA, CAB, and CBA. The number of possibilities is determined by factorial of number of members, which in here is $3! = 3\times2\times1 = 6$. In molecules that have symmetric fragments (group of atoms), there are identical cases among the possibilities; however, for bigger molecular fragments without symmetry (such as amino acids), the factorial rule applies.

Non-linear case scenarios

In organic molecules according to the carbon orbitals involved, the bond angles and hence the shape of the molecule will divert from simple linear example. In such cases, the number of possibilities will be affected by geometrical (*cis* and *trans*) and special (stereochemical-D and -L) arrangements and number of possibilities will increase dramatically.

How many molecules are there?

To obtain a better sense of the number crunching, an example is presented here. A group of scientists from the University of Bern has worked on the total number of possible organic molecules (Ruddigkeit, *et al*., 2012). The chemical space involved for molecules of up to 17 atoms of C, N, O, S, and halogens was calculated to be 166.4 billion entries. This forms the chemical universe database (GDB)-17, containing many drugs and lead compounds, and millions of isomers of known drugs. GDB-17 content, when compared to known molecules in PubChem, contains more nonaromatic heterocycles, stereoisomers, and scaffold types.

Among the possibilities in the chemical space, the bioactive ligands can be searched for by enumeration and subsequent virtual screening. It has been shown that almost all small molecules (>99.9%) have never been synthesized (Reymond and Awale, 2012), and more work has to be done for their preparation and laboratory testing. GDBs have been generated for prospective drug discovery purpose and are accessible.

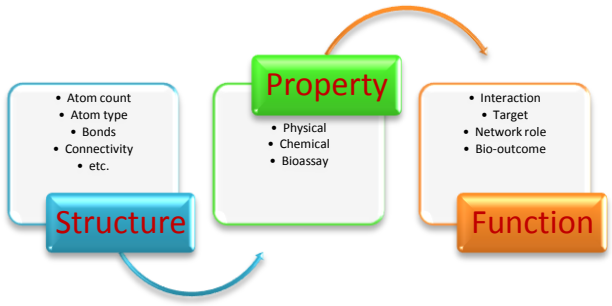
Higher up

Organic molecules are the center piece of the enumeration studies. They can be considered in groups as they are intact, or they can be split into smaller segments or fragments composing common moieties; then one or more of their properties or identification tags be converted into numbers to correlate with the function. In our recent study (Sardari *et al.*, 2016), utilizing combined pattern of methods such as fragment-based *de novo* design, scoring, similarity-based compound searching, and structure-based docking, led to introducing seven *in silico* designed compounds with antimycobacterial properties. Findings derived from antimycobacterial tests and MTT assay

indicated that 1-amino-4-(phenylamino) anthracene-9,10-dione and 5-fluoroindoline-2,3-dione have useful profile and are acceptable candidates for developing novel antimycobacterial drugs.

Closer to bio

Proteins are among the primary building blocks of the cell with multilevel complexity. Proteins are the main functional players in living organisms, and they can be annotated for their structure or function. As mentioned earlier, reaching an understanding for the function of biomolecules is always desired. A prediction for function can be performed according to homology-based approaches. A network of data also can fruitfully be impregnated with homology information to improve the predictive ability. The result is a network of proteins with homologous varieties linked across multiple species. Such ability is built in ProSNet algorithm by Wang *et al.* (2017). This integration of homology with molecular networks has been shown to substantially improve the predictive performance.



Proteins may act as biological receptors. G-protein-coupled receptors (GPCRs) are targets of many drugs due to their important functions such as neurotransmission, vision, immune response and so on. Numerical values have been used in building models for predicting (GPCRs) from amino acid sequence (Nie *et al.*, 2015). Such models were built with high accuracy based on fractal dimension, chaos game representation (CGR) and amino acid composition (AAC). The accuracy of 99.22% and correlation coefficient of 0.9845 was obtained for identifying GPCRs from non-GPCRs and accuracy of 85.73% to classify a GPCR into one of its five main subfamilies.

In summary, in order to study the structure and function of molecules and to tackle the problem of their huge number, numerical solutions can be applied. Other applications of enumeration in biomolecules and proteins include, but not limited to, tagging, brooding, homology modeling, improved BLAST algorithms, comparison matrices, mass spcectrometry analysis of proteins and finding the related peaks for the modified proteins and native ones in a mixture, pattern recognition for the interaction sites in proteins (protein-protein and protein-ligand) and analysis of interactome.

More details in:

A novel fractal approach for predicting G-protein-coupled receptors and their subfamilies with support vector machines. G Nie, Y Li, F Wang, S Wang, X Hu. Biomed Mater Eng; 2015. Vol. 1, Suppl 1: p. S1829-S836.

Exploring chemical space for drug discovery using the chemical universe database. JL Reymond, M Awale. ACS Chem Neurosci; 2012. Vol. 19;3, No. 9: p. 649-657.

Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. L Ruddigkeit, R van Deursen, LC Blum, JL Reymond. J Chem Inf Model; Vol. 26;52, No. 11: p. 2864-2875.

Fragment-based de novo design of antimycobacterial agents and in vitro potency evaluation. S Sardari, I Portugal, A ALKafri, D Moradi, G Ghavami. Comb Chem High Throughput Screen; 2016. Vol. 19, No. 3: p.238-245.

ProsNet: Integrating homology with molecular networks for protein function prediction. S Wang, M Qu, J Peng. Pac Symp Biocomput; 2017. Vol. 22: p. 27-38.