

SMILES OF A CHEMICAL STRUCTURE

SUBMITTED BY: ONEZA HASSAN ALVI

2023

REGISTRATION NUM: 04281913032

SUBMITTED TO: MISS. AFIFA NAVID

SYSTEM BIOLOGY (BIF-215)

NCB – QAU

# **QUESTIONS**

# **Question 1: What are SMILES of a chemical structure? Give 3 to 5 examples and briefly explain them.**

**Question 2:**

1. **How to convert a chemical structure in to SMILES and how to convert SMILES in to chemical structure?**
2. **Name 3 to 5 names of the tools or software used for the conversion**
3. **Attach Converted files**

**Question 3: In the context of Protein-Protein Interaction briefly define the following three terms.**

1. **P-value**
2. **E-value**
3. **Z-value**

# **ANSWERS**

# **Question 1: What are SMILES of a chemical structure? Give 3 to 5 examples and briefly explain them.**

Answer:

* SMILES: It stand for simplified molecular-input line-entry system.
* SMILES is the line notation for describing the chemical structures using the short ASCII strings. Their role comes in to play when the chemical structures are to be represented and processed using the computer systems.
* Important features of SMILES: SMILE strings can be imported using the molecule editors to convert them back in to 2-D drawings or 3-D chemical structures. Via the character- based approaches SMILES is used in various deep learning models. In the field of Drug Discovery SMILES also have a play in molecular property prediction, virtual screen etc. In Cheminformatics, to store and represent the chemical structure in databases the SMILES are used.
* SMILE examples:

1. Methane (CH4): C (i-e: the single carbon atom is represented by C and 4 H atoms are implicitly bonded to it.)
2. Hydrogen Per Oxide (H2O2): OO or O-O (i-e: the OO represent the two oxygen and 2 H atoms implicitly bonded to them.)
3. Ethyne (C2H2): C#C (i-e: the two C have triple bond)
4. Biphenyl (C12H10): C1CCCCC1-C2CCCCC2 (i-e: there are 2 benzene rings that have 2 carbon atoms. Rings are connected by single bond -. The first ring is C1CCCCC1 and 1 represent closure of sing. The second ring is C2CCCCC2 and 2 represent the ring closure.

**Question 2:**

1. **How to convert a chemical structure in to SMILES and how to convert SMILES in to chemical structure? Name 3 to 5 names of the tools or software used for the conversion**
2. **Attach Converted files**

Answer:

* **How to convert a chemical structure in to SMILES and how to convert SMILES in to chemical structure? Name 3 to 5 names of the tools or software used for the conversion?**

Chemical structures to SMILES and SMILES to Chemical structure: in order to do this manually first identify the atoms in the structure/SMILES and add its representation. Then represent the bonds and single and double or triple. Denote branches and represent the charges.

* SMILE notations:

1. Atoms: by their atomic symbols e-g; C represents Carbon atom
2. Bonds: single bond can be omitted or represented by – symbol. Double and triple bond by = and # respectively. Aromatic bonds are represented by :
3. Branches: represented by enclosing the branch in parenthesis. E-g: CC(C)C show it’s the three carbon atom chain and one carbon is attached as a branch to the second carbon in the chain.
4. Rings: they are represented by breaking any one bond in the ring and then add a number immediately after the atoms involved in the bond that was broken. E-g: C1CCCCC1 represents a benzene ring (6-C).
5. Charges: charges atoms are enclosed in square brackets. Charges are + and – symbols followed by magnitude. E-g: Ammonia cations are represented by [NH4+].

It is not necessary to use same tool or software to convert chemical structure to SMILES and SMILES to chemical structure. However, the use of same tools for conversions ensures accuracy.

Some of the software tools used for chemical structure and SMILES conversions include ChemDraw, Open Babel, RDKit, Marvin Sketch. Open Babel (efficient for chemical structure to SMILES conversions)

Some of the online tools used for chemical structure and SMILES conversions include PubChem Sketcher (efficient for SMILES to chemical structure convertion), MolView, Chem Spider, RCSB PDB Chemical Sketch Tool (efficient for chemical structure to SMILES convertion),

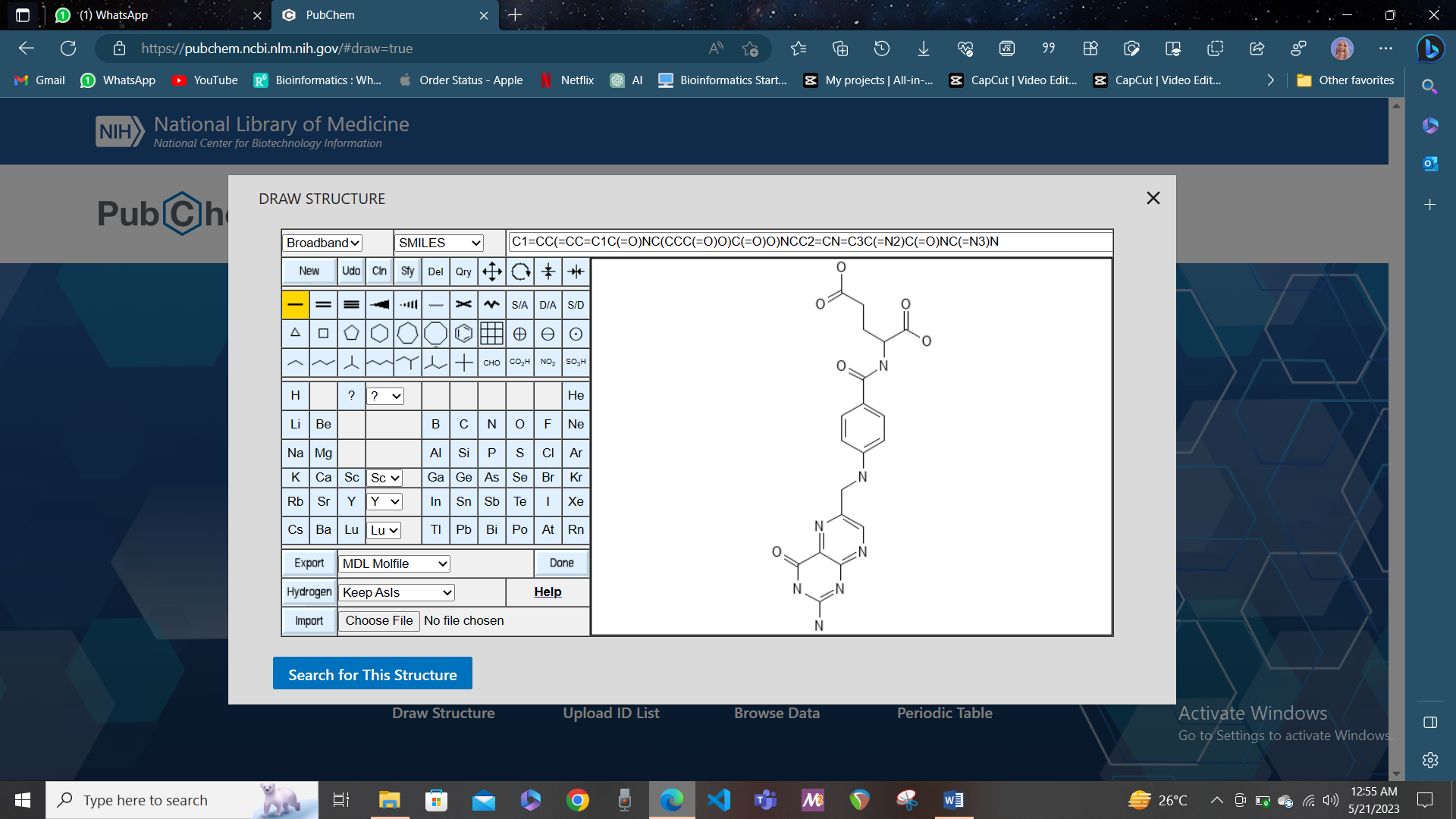
* **Attach Converted files**

**(The converted files are present in the folder, results are following)**

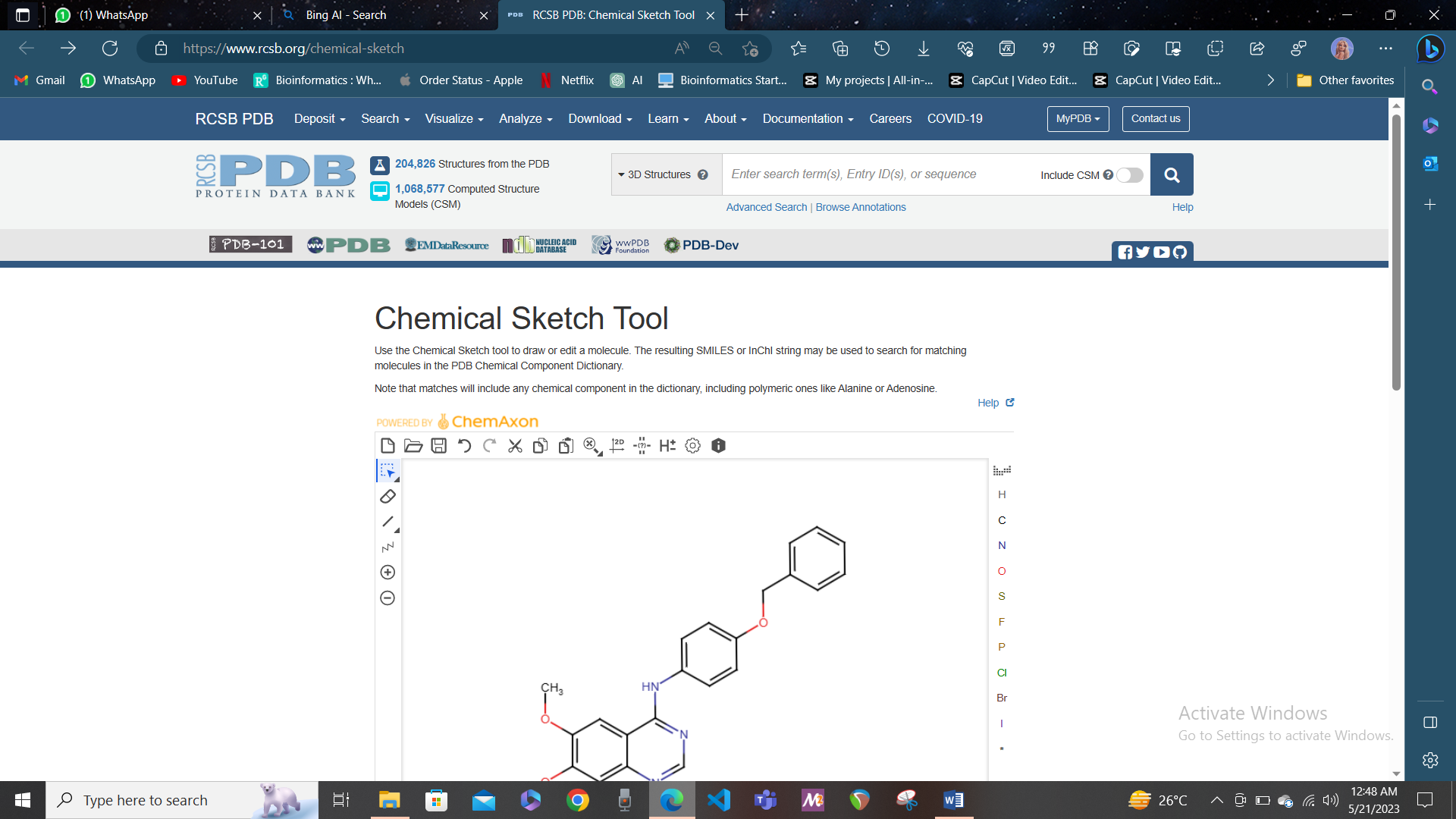
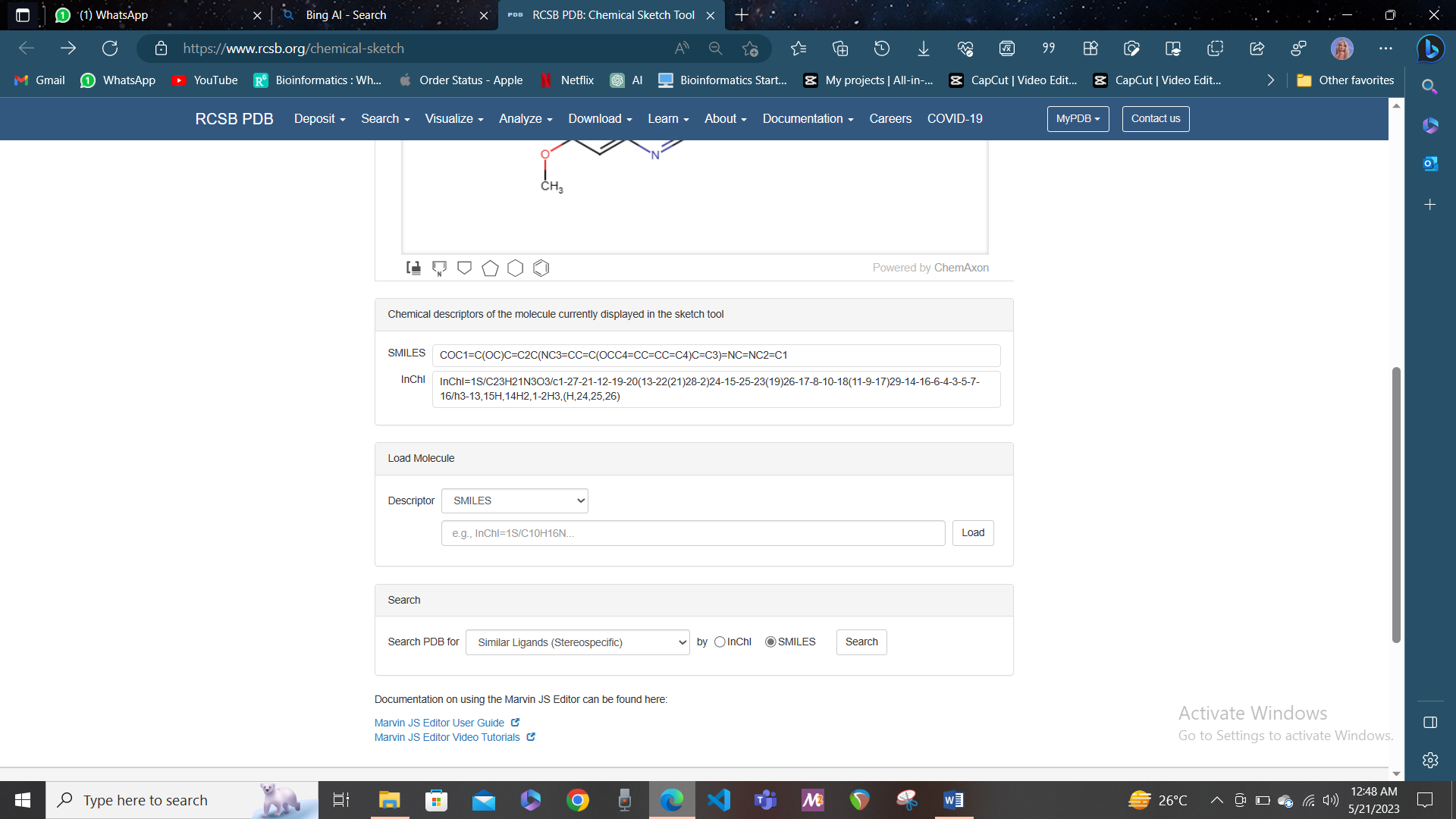
I have converted the folic acid canonical SMILE C1=CC(=CC=C1C(=O)NC(CCC(=O)O)C(=O)O)NCC2=CN=C3C(=N2)C(=O)NC(=N3)N

((2023), 2023)

I got the chemical structure using NCBI PubChem’s Sketcher and saved it in four formats i-e; pdf, png, InChl (.stdinchi), MDL Molfiles (.sdf)



I have converted EGFR/ErbB-2 Inhibitor Chemical structure that was stored in the MDL Molfiles (.sdf). The online web tool I used is RCSB PDB Chemical Sketch Tool and it generated the SMILES

The SIMLES of EGFR/ErbB-2 Inhibitor: COC1=C(OC)C=C2C(NC3=CC=C(OCC4=CC=CC=C4)C=C3)=NC=NC2=C1

**Question 3: In the context of Protein-Protein Interaction briefly define the following three terms.**

1. **P-value**
2. **E-value**
3. **Z-value**

Answer:

1. P-value: P value is used to detect and recognize the statistical significance of the protein-protein interaction that is observed. It is a measure of probability that the interaction between proteins occurs by chance. A low p values indicates that the observed interaction are unlikely to have taken place by chance and thus they have great significance, strong interaction. It mean that these interactions with low p value (for example <0.05) have biological significance and they are not random occurrence.
2. E-value: it is commonly used in sequence alignment tool (e-g: BLAST) to recognize the random matches that are found against the query in the database. In case of protein-protein interaction it tells the signification of the sequence similarity of the proteins in the interactions. Low E-value indicates that the significance of similarity is high and the matching is not on random occurrence, thus statistically significant.
3. Z-score: is depicts the number of the standard deviation the data point is from the mean. It recognize how far the data points is from the mean distribution and detects the outlier. However it is not commonly used in the protein-protein interactions analysis the other two parameters are more significant.