SciForum
Mol2Net

Pro-ChInt: Machine Learning Methods for Identifying Dual-/Multi- Protein Chains Interactions with Python

Yong Liu [*a]

[a] Information and Communication Technologies Department, Faculty of Computer Science, University of A Coruna, 15071 A Coruña, Spain

* Correspondent author: Yong Liu; e-mail: y.liu86@outlook.com; Tel.: +34 981 167 000; Fax: +34 981 167 160.

Published: 7 December 2015

Abstract: In nature, protein chain interactions (Pro-ChInt) of single- / multi-protein, a common but complex system, refer to physical contacts established between two or more protein chains depending on the amino acid sequences, which contains tremendous information. Decoding amino acid sequence information of protein using complex networks or graphs of the peptides is a grateful solution to discover the communication information between different Pro-ChInt. We first constructed some python codes to directly download the specify protein sequences from the RCSB protein data bank (PDB). Then, we changed the FASTA format to S2SNet format to calculate the embedded / non-embedded parameters of protein chains according to the star graph topological indices of peptide sequences. Meanwhile, we numbered all protein chains, then used the chain numbers to get a random number for a given set of chain number or case number used for each protein. Then, we replaced all the random numbers with the corresponding parameters of each protein chain calculated with S2SNet application. After that, a machine learning classification model was constructed based on the combinatorial / combining interaction of different chains. This new method can be used to identify two or more protein chain interactions combined with machine learning technique.

Keywords: Machine Learning, Protein Interaction, Protein Sequence, Python Scripts, Classification Model.

1. Introduction

Proteins are the main components of the biological metabolic pathways in living organisms. In nature, it could be one individual chain, or more than two chains to constitute a functional complex organic whole. Generally, the communications among different protein chains are very complicated, how to decode the communicational "language" is an important research topic in current chemoinformatics, bioinformatics, and pharmaceutics.

The biological systems are very complicated, therefore, a lot of scientists try to account for the biological complex problems with the techniques of genomics, transcriptomics and proteomics. However, proteomics are more complicated than genomics as genome is generally constant, whereas the proteome differs lie on cell and time. Proteins are subjected to a wide variety of chemical modifications after translation. It called as post-translational modifications, such as phosphorylation, ubiquitination, methylation, oxidation, *etc*.

Well understood of protein molecular information is helpful to disease control or prevention. This is because structure decides function for proteins. Whereas, proteins are the "practitioner" directly participating in the complex biological life cycle. In nature, protein – protein interactions refer to the physical contacts established between two or more proteins by the electrostatic forces and/or biochemical events. Whereas, the functional domains are generally formed by two or more protein chains but not only one chain or one protein. Decoding amino acid sequence information of protein, using complex networks or graphs of the peptide, is a grateful solution to uncover protein chain – chain interaction (Pro-ChInt).

Some sequence to structure graphs are used to calculate the numeric descriptors of molecular structure, for instance, MARCH-INSIDE [1] and S2SNet [2, 3]. These tools can transform the characters and numeric sequences into Star network graph. And then to calculate Star Graph Topological Indices.

2. Results and Discussion

In present work, we first searched the target PDB-ID with some special performance, and save all PDB-ID in a text file. Then we got all the FASTA profile of protein chain by using the python module "urllib2". We transformed the FASTA to S2SNet format, some examples of FASTA and S2SNet profile was presented in **figure 1**. S2SNet format is easier for further work.

```
>5AKA:0|PDBID|CHAIN|SEQUENCE
AVQQNKPTRSKRGMRRSHDALTAVTSLSVDKTSGE
>5AKA:1|PDBID|CHAIN|SEQUENCE
AKGIREKIKLVSSAGTGHFYTTTKNKRTKPEKLEL
>5AKA:2|PDBID|CHAIN|SEQUENCE
MKRTFQPSVLKRNRSHGFRARMATKNGRQVLARRR
>5AKA:3|PDBID|CHAIN|SEQUENCE
PKIKTVRGAAKRFKKTGKGGFKHKHANLRHILTKK
>5AKA:4|PDBID|CHAIN|SEQUENCE
MKVRASVKKLCRNCKIVKRDGVIRVICSAEPKHKQ
>5AKA:5|PDBID|CHAIN|SEQUENCE
GFDLNDFLEQLRQMKNMGGMASLMGKLPGMGQIPD
MQVQDVNRLLKQFDDMQRMMKKMKKGGMA
>5AKA:6|PDBID|CHAIN|SEQUENCE
```

```
PDBID  Chain  Seq.
1QF6   A      MPVITLPDGSQRHYDHAVSP
2J28   0      AVQQNKPTRSKRGMRRSHDA
2J28   1      AKGIREKIKLVSSAGTGHFY
2J28   2      MKRTFQPSVLKRNRSHGFRA
2J28   3      PKIKTVRGAAKRFKKTGKGG
2J28   4      MKVRASVKKLCRNCKIVKRD
2J28   7      LGFPINFLTLYVTVQHKK
2J28   9      FDNLTDRLSRTLRNISGRGR
2J28   C      VKCKPTSPGRRHVVKVVNPE
2J28   D      MIGLVGKKVGMTRIFTEDGV
2J28   E      MELVLKDAQSALTVSETTFG
2J28   F      AKLHDYYKDFVVKKLMTEFN
```

FASTA format S2SNet format

Figure 1. The FASTA and S2SNet profiles of some protein chain

After to get the S2SNet format, the TIs parameters of each protein chain was calculated by S2SNet application. In here, we also can use others methods to calculate the molecular descriptors for protein sequences. For instance, we are trying to divide all the amino acids into four different types, polar or non-polar, charged or uncharged amino acids. We can count the number of polar-polar amino acids, polar-x-polar amino acids, and polar-x…x-polar amino acids. Or other types of connection between the different amino acids. However, this part of work, we have not yet finished. So in here, we used S2SNet, one of previous work in our group.

On the other hand, we numbered all the chain in a given file, and to select the corresponding numbers of each protein to run a random selection among the given chain numbers (n). For example, the first protein has 9 chains, these chains have the number from 1 to 9. We let users to put the number (m-fold), we can get the random cases = n × m. However, we have to remove the duplicates before we get all the final cases. The more important part of present work is to define how many of chain will be assigned to run the interaction between one to others. For example, with our new codes, we can perform the interaction among two or more (depending on the users, **Figure 2**).

```
3782 3710 3761
3728 3737 3757
3726 3748 3773
3756 3714 3770
3713 3720 3760
3790 3708 3724
3783 3788 3739
3780 3723 3751
3782 3787 3735
3749 3726 3723
3752 3745 3768
3764 3738 3757
3744 3720 3758
3791 3771 3749
3752 3732 3715
```

3 chains

```
5351 5485 5486 5485 5305 5406
5297 5359 5306 5434 5395 5314
5356 5305 5387 5466 5402 5465
5502 5404 5455 5459 5336 5467
5466 5392 5402 5317 5333 5306
5412 5317 5323 5297 5427 5326
5499 5394 5333 5419 5418 5464
5364 5430 5487 5490 5322 5361
5375 5438 5309 5324 5401 5408
5339 5427 5408 5331 5477 5450
5381 5368 5483 5372 5387 5349
5473 5454 5384 5349 5358 5305
5455 5406 5319 5419 5499 5493
```

6 chains

```
886 918 922 964 888 915 872
889 881 908 919 938 907 950
931 951 918 912 915 923 894
904 875 912 885 931 963 881
881 889 940 963 913 883 888
917 906 895 873 909 959 913
939 896 932 961 897 888 886
928 930 873 956 922 904 931
952 962 916 902 947 911 921
901 884 884 931 922 876 886
947 945 878 894 958 891 895
903 886 940 888 937 886 873
877 928 901 880 923 908 953
924 884 888 949 881 937 870
927 964 926 879 950 962 884
```

7 chains

Figure 2. The examples of random numbers for the chain-chain interaction

In addition, each random number refers to the corresponding chain sequence, and each sequence would be calculated into 42 TIs. If all the sequences (numbers) in the combination are from the same protein, we defined this case as the "positive" or "1", whereas, if not all numbers from the same protein, we consider this case as the "negative" or "0".

After that, we obtained and calculated each combination character based on the average values of each combination (42 TIs average values). Using this data to run a classification model to identify if there are interactions among the two or more chains.

3. Material and Methods

All codes were programmed in the platform of PyCharm 3.5 version under the environmental of python 2.7 version. There are different steps to establish Pro-ChInt. They include to obtain the target protein chains sequence, change FASTA to S2SNet format and calculate S2SNet star graph topological indices, to get the serial random number, etc.

3.1. Download FASTA files

First step, we programed the codes in python to download the FASTA file from the protein data bank (PDB) according to the PDB-ID (serial number). In this part, we used urllib2 module to corresponding website of specify protein ID. The codes presented as following:

```
with open("pdblist.txt", 'r') as fout:
    pdb_list = fout.read()
    pdb_list = pdb_list.strip()
    for pdb in pdb_list.split('\n'):
        sequence_url =
'http://www.rcsb.org/pdb/files/fasta.txt?structureIdList
=' + pdb.strip()
        response = urllib2.urlopen(sequence_url)
        pdb_text = response.read()
        pdb_text_str = str(pdb_text)
        inFasta_text = pdb_text_str
        inFasta.write(inFasta_text)
        inFasta.write('\n')
```

3.2. Transform FASTA to S2SNet format

In this part, we change the FASTA format to S2SNet format for the further calculation. The details python codes presented in GitHub: https://github.com/muntisa/pyS2SNet/blob/master/pyScripts/3_S2SNetFilterByPDBchains.py.

```
foutFile = open("S2SNetchains.txt",'w')
for sline in linesFASTA:
    ilines += 1
    if sline[0] == '>':
        PDBfasta=sline[1:5]
        ChainFasta = sline[6:8]
        if ChainFasta[1] == '|':
            ChainFasta = ChainFasta[:-1]
        else:
            ChainFasta= ChainFasta
        if ilines !=1:
            Seq = Seq + "\n"
        Seq =
Seq+str(PDBfasta)+"\t"+str(ChainFasta)+"\t"
    else:
        Seq = Seq+sline[:-1]
foutFile.write(Seq)
foutFile.close()
```

3.3. Calculate the S2SNet topological indices

We obtained 42 topological indices (TIs) for each protein chain, calculated by S2SNet star graph. There are two types of TIs (Embedded and non-Embedded indices). Each one has 21 TIs. Like Shannon entropies, connectivitymatrices, Harary number, Wiener index, Gutman topological index with different power [4]. S2SNet are widely used in obtaining the molecular information of protein [5].

3.4. Get the random number matrix

In this part, we first numbered all the protein chains according to the order of all chain appeared in the PDB chain file. Each protein chain has the only special number. For example, the first protein has 9 chains (n), these chains have the number from 1 to 9, but for the second protein, if it has 15 chains, the number of these chains are from 10 to 24, and so on. Then, we used two codes to let the users to input the chain number, how many chains (Maximum) will be accounted for Pro-ChInt. Meantime, we let users to put the number (m-fold), we can get the random cases = n × m. In final, we remove all the replicated cases.

```
# input the chain number and the case numbers for each
protein
ChainNumber = raw_input("Input chain number(k>=2): ")
RowNumber = raw_input("Input each protein chain Multiple
(m): ")
```

3.5. Classification modeling

In this step, we replace the random number with the 42 TIs of corresponding protein chain sequence. For one combination, if all the chains are from the same original protein, we consider this combination has the chain-chain interaction (Pro-ChInt) set as "1 or positive". If the combination is from different protein, we consider this combination has no chain-chain interaction, Pro-ChInt, set as "0 or negative".

For each combination, we calculate the average value of each parameter in 42 TIs of S2SNet Star Graph. After that, we can use Weka to obtain the best classification model depending on the combination mentioned previous.

4. Conclusions

This short communication is presenting some original python codes for identify the protein chain – chain interactions lie on the S2SNet Star Graph Topological Indices. The ideas of this work are on account of molecular descriptors obtained from Star Graphs. Then to use Machine Learning methods running in Weka to search for the best classification model. We can explain the protein chain – chain interaction based on the molecular information of protein sequences.

Acknowledgments

The authors acknowledge the support provided by the Galician Network of Drugs R+D REGID (Xunta de Galicia R2014/025). This work was

partially supported by the Galician Network for Colorectal Cancer Research (Red Gallega de Cáncer Colorrectal - REGICC, Ref.: CN 2012/217), Institute for Biomedical Informatics of A Coruña (INIBIC), and Center for Research of Information and Communication Technologies (CITIC).

Conflicts of Interest

The authors declare no conflict of interest.

References

1. González-Díaz, H.; Torres-Gomez, L. A.; Guevara, Y.; Almeida, M. S.; Molina, R.; Castanedo, N.; Santana, L.; Uriarte, E. Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer-aided molecular design III: 2.5D indices for the discovery of antibacterials. *J Mol Model* **2005,** *11* (2), 116-23.
2. Munteanu, C. R.; Gonzáles-Díaz, H. *S2SNet - Sequence to Star Network, Reg. No. 03 / 2008 / 1338, Santiago de Compostela, Spain*, Santiago de Compostela, Spain, 2008.
3. Munteanu, C. R.; Magalhaes, A. L.; Duardo-Sanchez, A.; Pazos, A.; Gonzalez-Diaz, H. S2SNet: A Tool for Transforming Characters and Numeric Sequences into Star Network Topological Indices in Chemoinformatics, Bioinformatics, Biomedical, and Social-Legal Sciences. *Current Bioinformatics* **2013,** *8* (4), 429-437.
4. Fernández-Blanco, E.; Aguiar-Pulido, V.; Munteanu, C. R.; Dorado, J. Random Forest classification based on star graph topological indices for antioxidant proteins. *Journal of Theoretical Biology* **2013,** *317*, 331-337.
5. Liu, Y.; Munteanu, C. R.; Fernández Blanco, E.; Tan, Z.; Santos del Riego, A.; Pazos, A. Prediction of Nucleotide Binding Peptides Using Star Graph Topological Indices. *Molecular Informatics* **2015,** *34* (11-12), 736-741.