Comparison of Protein-Protein Interaction from Geometry and Biochemistry View with Computation-Driven Data

Shree Sundari Devi D/O S[1] **Kwoh Chee Keong**[1] **Prasanna R Kolatkar**[2]
[1] *School of Computer Engineering, Nanyang Technological University, Singapore*
[2]*Genome Institute of Singapore, Singapore*
Email : shreedev@singnet.com.sg, asckkwoh@ntu.edu.sg, kolatkarp@gis.a-star.edu.sg

ABSTRACT: In this paper, we present a tool to calculate the distribution of amino acid contacts in proteins as well as in protein domains. The proteins are grouped according to the classification by Yanay Ofran and Burkhard Rost[1]. In addition, a protein's distribution was compared with that of proteins in the same group as well as the entire collection of proteins across all groups. With these statistics, biologists can pick out proteins which have characteristics that defer from the norm.

1 INTRODUCTION

1.1 Contact Residue Pairs

A residue pair is defined to be in contact if the distance of the closest of their respective atoms is less than 6 angstroms and they are separated by three or more residues. A contact residue pair may originate from the same polypeptide chain of the protein, or from different chains and are referred as internal chain contact residue pairs and external chain contact residue pairs respectively

1.2 Classification of Proteins

Previous study has been done by Yanay Ofran and Burkhard Rost[1] to classify proteins into six interfaces or groups, according to their type of interactions, which include interactions within the same domain of a polypeptide chain, interactions across different domains of a polypeptide chain, transiently interacting proteins, permanently interacting proteins and so on. These groups and their properties are shown in Table 1.

Group / Interface	Property Of Group / Interface
Intra – Domain	Interfaces within one structural domain and in the same chain of the protein
Domain – Domain	Interfaces between different domains within one chain
Homo – Obligomer	Interfaces between permanently interacting identical chains
Homo – Complex	Interfaces between transiently interacting identical protein chains
Hetero – Obligomer	Interfaces between permanently interacting different protein chains
Hetero – Complex	Interfaces between different transiently interacting protein chains

Table 1: Different Classifications of Proteins and their properties as concluded by Ofran and Rost[1]

It was concluded by Ofran and Rost[1] that each group had a different structural association between residues. They also discovered significant differences in amino acid composition and residue–residue preferences between interactions of residues in the proteins belonging to the six groups. The differences between the six groups were so vast that they were able to statistically predict the group each of a set of 1000 residues, belonged to using amino acid composition alone with an accuracy of 63 to 100 percent. Some of their findings and a description of how the results gathered from this project validate or invalidate their conclusions are discussed in Section 3.2.1.

2 METHODOLOGY

A statistical approach was used to represent the distribution of amino acid contacts in proteins and protein domains. In lieu of the above, 6 sets of statistics were calculated and a mathematical symbol was used to represent each statistic.

The naming conventions in the mathematical equations are defined as follows:

Let Y represent a contact between two residues.

$Y_{a_i a_j c_k c_l}$ represents a contact between amino acid *i* from chain *k* and amino acid *j* from chain *l*.

The set of amino acids, *a*, is represented as follows:

$$a \in \{ALA, ARG, ASN, ASP, CYS, GLN, GLU, GLY, HIS, ILE, \\ LEU, LYS, MET, PHE, PRO, SER, THR, TRP, TYR, VAL\}$$

The set of protein groups, *g*, is represented as follows:

$$g \in \{Hetero\,Complex, Homo\,Complex, Hetero\,Obligomer, \\ Hetero\,Complex, Domain - Domain, Intra\,Domain\}$$

Let c represent the set of chain ids present in protein, *p*. c_k represents a particular chain id present in c.

Let N represent the total number of proteins being analysed. N_{g_x} represents the number of proteins in group *x*.

The sets of statistics generated are described in Section 2.1.

2.1 Calculation of Statistics

2.1.1 Probabilities of Contacts

The distribution of amino acid contacts in proteins were determined by calculating the statistics representing the probability of all possible amino acid – amino acid contacts in a protein. The statistics calculated were overall probability, overall conditional probability, internal chain probability, internal chain conditional probability, external chain probability and external chain conditional probability. The mathematical equations representing each set of statistics are shown in Table 2.

	Probability	Conditional Probability
Overall	$P(Y_{a_ia_j}) = \frac{\lvert Y_{a_ia_j}\rvert}{\sum_{i=1}^{20}\sum_{j=1}^{20}\lvert Y_{a_ia_j}\rvert}$	$P(Y_{a_ia_j} \mid Y_{a_i}) = \frac{\lvert Y_{a_ia_j}\rvert}{\sum_{j=1}^{20}\lvert Y_{a_ia_j}\rvert}$
Internal Chain	$P(Y_{a_ia_jc_kc_k}) = \frac{\lvert Y_{a_ia_jc_kc_k}\rvert}{\sum_{i=1}^{20}\sum_{j=1}^{20}\lvert Y_{a_ia_jc_kc_k}\rvert}$	$P(Y_{a_ia_jc_kc_k}) = \frac{\lvert Y_{a_ia_jc_kc_k}\rvert}{\sum_{j=1}^{20}\lvert Y_{a_ia_jc_kc_k}\rvert}$
External Chain (where l ≠ k)	$P(Y_{a_1a_{20}c_kc_l}) = \frac{\lvert Y_{a_1a_{20}c_kc_l}\rvert}{\sum_{i=1}^{20}\sum_{j=1}^{20}\lvert Y_{a_ia_jc_kc_l}\rvert}$	$P(Y_{a_1a_{20}c_kc_l} \mid Y_{a_1c_kc_l}) = \frac{\lvert Y_{a_1a_{20}c_kc_l}\rvert}{\sum_{j=1}^{20}\lvert Y_{a_1a_jc_kc_l}\rvert}$

Table 2: Mathematical equations representing each type of statistics calculated

The overall probability of a particular amino acid – amino acid contact refers to the number of such amino acid – amino acid contact pairs in the protein, with respect to all contact pair residues in the protein. Similarly, the internal chain probability of a particular amino acid – acid contact refers to the number of such amino acid – amino acid contact pairs from the same chain, with respect to all contact pair residues in that particular chain and the external chain probability refers to the number of such contacts from different chains with respect to all contacts from different chains.

The overall conditional probability of a particular amino acid – amino acid contact refers to the number of such amino acid – amino acid contact pairs in the protein, with respect to all contact pair residues of the first amino acid. The internal chain conditional probability of a particular amino acid – amino acid contact refers to the number of such amino acid – amino acid contact pairs within the same chain, with respect to all contact residue pairs of the first amino acid in that particular chain and the external chain conditional probability refers to number of such contacts from different chains with respect to all contacts of the first amino acid, from different chains.

These sets of statistics were calculated for each protein, following which, the group average statistics, representing the mean of all proteins in a group and the overall average statistics, representing the mean of the entire set of proteins, across all groups were calculated.

2.2.2 Group and Overall Average

The group and overall average is a measure of the group and overall norm. The group average probability was calculated as follows:

Average Probability (Amino acid *i* in contact with amino acid *j* in group *x*), $\overline{P(Y_{a_ia_j})_{g_x}}$

$$= \frac{\sum_{p=1}^{N_{gx}} P(Y_{a_ia_j})}{N_{gx}}$$

Similarly, the overall average probability of particular amino acid – amino acid contact was calculated as follows:

Average Probability (Amino acid *i* in contact with amino acid *j* in all proteins), $\overline{P(Y_{a_ia_j})}$

$$= \frac{\sum_{x} \overline{P(Y_{a_ia_j})_{gx}}}{|g|}$$

2.2.3 Probabilities of Contacts in Protein Domains

The type and location of domains in a particular protein was obtained from a text file, downloaded for the PFam FTP Server[3].

The same set of equations was used to determine the distribution of amino acid contacts in protein domains. However, the probabilities were calculated with respect to all contact pairs within the domain, rather than in the whole protein. In other words, only contacts within the residue range defined by the domain were taken into account.

The equations used to calculated the probability and conditional probability of amino acid contacts in protein domains are as follows:

Probability (Amino acid *i* in contact with amino acid *j* in domain *m*),

$$= \frac{\text{Number of residues in which amino acid } i \text{ and with amino acid } j \text{ are in contact in domain } m, \; P\left(Y_{d_m a_i a_j}\right)}{\text{Total number of contact residues in domain } m \text{ in protein}}$$

$$= \frac{\left|Y_{d_m a_i a_j}\right|}{\sum_{i=1}^{i=20} \sum_{j=1}^{j=20} \left|Y_{d_m a_i a_j}\right|}$$

Conditional Probability(Amino acid *i* in contact with amino acid *j*),

$$= \frac{\text{Number of residues in which amino acid } i \text{ and with amino acid } j \text{ are in contact in domain } m, \; P\left(Y_{d_m a_i a_j} \mid Y_{d_m a_i}\right)}{\text{Total number of residues in contact with amino acid } i \text{ in domain } m}$$

$$= \frac{\left|Y_{d_m a_i a_j}\right|}{\sum_{j=1}^{j=20} \left|Y_{d_m a_i a_j}\right|}$$

2.2.4 Z-Score

A protein's amino acid contacts distribution was compared with of the group and overall average and it's deviation from the average was measured by the calculation of its z-score. The z-score is a measure of how far and in what direction, an item deviates from the mean of the distribution. In this case, the z-score measures how far a particular protein deviates from the mean of the group, as well as the mean of the entire collection of proteins.

The formula for converting any of the six probabilities discussed in Section 2.2.1 into its corresponding Z-Score are shown in Equations 1 and 2.

Z-Score of the probability of amino acid i contacting with amino acid j, with respect to its group mean, $Z_{P(Y_{a_i a_j c_k c_l})_{gx}}$

$$= \frac{P\left(Y_{a_i a_j c_k c_l}\right) - \overline{P\left(Y_{a_i a_j c_k c_l}\right)}_{gx}}{\sigma_{P(Y_{a_i a_j c_k c_l})}} \qquad (1)$$

Z-Score of the probability of amino acid I contacting with amino acid j, with respect to the overall mean, $Z_{P(Y_{a_i a_j c_k c_l})o}$

$$= \frac{P\left(Y_{a_i a_j c_k c_l}\right) - \overline{P\left(Y_{a_i a_j c_k c_l}\right)}}{\sigma_{P(Y_{a_i a_j c_k c_l})}} \qquad (2)$$

3 RESULTS

3.1 Average Distribution of Amino Acid Contacts in the Six Groups

On the whole, all the six groups have similar distributions, with a peak at the Leucine – Leucine contact. It was also observed that the other high probability values were as a result of other Leucine contacts, such as Leucine – Valine, Leucine – Isoleucine and Leucine – Alanine. Valine - Valine contacts also occur with a high probability in all six groups. When the probabilities of all contacts were arranged in descending order, all Leucine contacts fall in the top 10% (i.e. the top 40 with respect to the 400 possible types of contacts), across all the six groups.

It was also observed that Hetero – Obligomers, in particular, have a high probability of Cysteine – Cysteine contacts. Cysteine – Cysteine contacts occur in the above group with a probability of 0.0111, which is the second highest probability of contact in that group, after the Leucine – Leucine contact.

Leucine contacts occur with the highest probability and Hetero-Obligomers have a high probability of Cysteine – Cysteine internal chain contacts as well. Leucine contacts also occur with a high probability in external chain contacts, with the exception of Intra-Domain proteins. Intra-Domain proteins are categorized as those which have prominent internal chain contacts. Therefore, the probability of external chain contacts in proteins in the group Intra-Domain, is very small.

3.2 Distribution of Amino Acid Contacts in Terms of Chemical Bonds

This section provides a brief discussion on the distribution of amino contacts in terms of the chemical bonds formed.

It was observed that all protein groups are rich in hydrophobic interactions. Contacts which may give rise to hydrophobic interaction occur with the highest probability in all groups. Leucine is non-polar and therefore water-hating. The high percentage of Leucine contacts gives rise to the hydrophobic interaction.

3.2.1 Comparison with the Findings of Ofran and Rost[1]

The following was concluded by Ofran and Rost[1]:

1. Homo complexes are depleted in salt bridges, but rich in contacts between identical residues.
2. Cysteine bridges occurred more than expected in all groups.
3. Salt bridges are common in all groups, with the exception of Homo Complexes.

The results obtained comply with Ofran and Rost's[1] findings to a certain extent. However, it was observed that

Homo Complexes do not seem to have a depletion of salt bridges, when compared to other groups. Salt bridges, on the whole, are common among all groups. With the exception of the above, the results obtained does comply with the conclusions drawn by Ofran and Rost[1].

3.3 Proteins with High Z-Score Values

The bulk of proteins with high z-scores are those which very few (less than 100) contacts. These proteins, having very few contacts will naturally have z-scores when compared to the group and overall averages. However, there are some larger proteins with high z-scores as well. Protein id 2SIV, with 839 contacts, has a z-score of 15.285 when its probability of the Glutamine – Glutamine contact was compared with its group norm and a z-score of 15.297 when compared to the overall norm. This is because the Glutamine – Glutamine contact in 2SIV occurs with a very high percentage. Protein 1NPO, with 799 contacts, has a z-score of 10.261 when its probability of the Cysteine – Glutamine contact was compared with its group norm and a z-score of 10.011, when compared to the overall norm. In 1NPO, 12 out of the possible 20 types of Cysteine contacts have group and overall z-score values between 3 and 11. This shows 1NPO is very rich in Cysteine contacts in comparison to its group (Homo – Obligomers) and the overall average. The above-mentioned proteins and some other bigger proteins which have high z-scores are summarized in Table 3.

Protein ID	Group	Type of Contact	Group Z-Score	Overall Z-Score
2SIV	Hetero Complexes	Glu – Glu	15.285	15.297
1NPO	Homo Obligomers	Cys – Glu	10.261	10.011
1CNO	Homo Obligomers	Ala – Glu	10.44596	10.4457
1OVO	Homo Obligomers	Cys – Asp	13.207	12.914
1EZG	Homo Obligomers	Cys – Thr	18.29628	18.1919
1EZG	Homo Obligomers	Thr – Thr	14.52797	14.5375
1FD3	Homo Obligomers	Cys – Ile	14.64431	14.6298

Table 3: List of proteins with a large number of contact pairs and high z-score values

3.4 Interaction between Polar and Non-Polar Residues

It was observed there are contacts between polar and non-polar residues resulting from either electrostatic atoms or hydrophobic portions coming together. A summary of contacts of the above type which occurred with a high percentage in each group is shown in Table 4.

Group	Type of Contact
Hetero Complex	Leucine – Serine
	Leucine – Threonine
	Threonine – Valine
Homo Complex	Leucine – Serine
	Glutamine – Alanine
Hetero Obligomer	Leucine – Threonine
	Alanine – Serine
Homo Obligomer	Leucine – Threonine
	Leucine – Serine
	Leucine – Tyrosine
Domain – Domain	Leucine – Threonine
	Leucine – Serine
	Leucine – Tyrosine
	Threoine – Valine
	Leucine – Glutamic Acid
Inter – Domain	Leucine – Threonine
	Leucine – Tyrosine
	Leucine – Lysine

Table 4: List of contacts between polar and non-polar residues, which occurred with a high probability in each group

The Leucine – Threonine contacts occurred at a high percentage in every group. However, Leucine, being a non-polar molecule is hydrophobic and Threonine which is a hydroxyl and therefore polar, is hydrophilic. The same can be said for Leucine – Serine, Leucine – Tyrosine, Valine – Threonine and Alanine – Serine contacts, which involve a contact between a non-polar hydrophobic molecule and a polar, hydrophilic hydroxyl. Leucine and Valine have high hydrophobicity values of 0.943 and 0.825, whereas hydroxyls Serine and Threonine have hydrophobicity values of 0.359 and 0.450 respectively.

The Glutamine – Alanine contact which occurred with a high percentage in Homo Complexes is a contact between a polar, hydrophilic amide and a non – polar hydrophobic molecule. The Leucine – Glutamic acid contact in the group Domain – Domain is a contact between a polar, hydrophilic acid and a non-polar hydrophobic molecule. Moreover, the hydrophobicity value of Glutamic acid is 0.043, which very much lower than that of Leucine.

3.5 Discussion: Distribution of Amino Acid Contacts in Protein Complex, Oct4/Sox2

This section provides a brief description on the amino acid distribution of the protein, Oct4/Sox2. Oct4/Sox2, is a polymerase which starts RNA transcription.

Oct4/Sox2 is a protein complex comprising transcription factors, Oct4 and Sox2. As the structural information of the above protein complex is not available, the structure of the complex, Oct1/Sox2, which is a close homologue of Oct4/Sox2, was used to determine the amino acid distribution.

The objective was to compare the distribution of the complex with that of Oct1 and Sox2 individually. A search in the Protein Data Bank[2] website (http://www.rcsb.org/pdb) for Oct1, Sox2 and the Oct1 Sox2 complex, resulted in the PDB ids 1OCT, 1GT0 and 1O4X. 1OCT represents the structure of the compound Oct1. 1GT0 is actually another complex comprising of Sox2 and three other molecules. As the structure of Sox2 alone, was not available, 1GT0 was the next closest match. 1O4X represents the structure of a complex comprising Oct1, Sox2 and two other molecules. In lieu of this, comparing the above three proteins based on the amino acid contact distribution in the whole protein will not be accurate. Therefore, the comparison was done based on the distribution of amino acid contacts in the domains of the three proteins. The domains present in all three proteins are summarized in Table 5.

Protein ID	Domains
1OCT	Homeobox (Chain C: 102 – 158)
	Pou (Chain C: 5 – 75)
1GT0 (in the Sox2 molecule)	HMG_Box (Chain D: 3 – 71)
1O4X	Homeobox (Chain A: 110 – 161)
	Pou (Chain A: 5 – 79)
	HMG_Box (Chain B: 208 – 276)

Table 5: List of domains present in proteins 1OCT, 1GT0 and 1O4X

From Table 5 it can be observed that the domains Homeobox and Pou are present in 1OCT and the Sox2 molecule gives rise to the HMG_Box domain in 1GT0. 1O4X, being the complex of Oct1 and Sox2, contains all the three domains.

As per our results, the number of contacts in each domain, in the three proteins is shown in Table 6.

Protein ID	Domain	Number of Contacts
1OCT	Homeobox	284
	Pou	454
1GT0	HMG_Box	348
1O4X	Homeobox	384
	Pou	454
	HMG_Box	348

Table 6: Number of contacts in the domains present in 1OCT, 1GT0 and 1O4X

Table 6 shows that with the exception of the domain, Homeobox, the number of contacts in the same domain is equal. It was also observed that the overall distribution of contacts in the domains in the protein complex, 1O4X is similar to that of 1OCT and 1GT0, which contain Oct1 and Sox2 alone.

The Isoleucine – Isoleucine and Isoleucine – Leucine contacts occur with the highest percentage in Homeobox domain in both 1OCT and protein complex 1O4X. The peaks in the distribution of the Pou domain correspond to the Leucine – Leucine and Leucine – Phenylalanine contacts. There are no distinct peaks in the distribution of the HMG_Box domain with many contacts occurring with a probability of around 0.1.

4 CONCLUSION

The distribution of amino acid contacts in proteins was calculated and the distribution was compared to that of the group and overall average to determine the extent of deviation. In addition, the distribution of amino acid contacts in protein domains was also calculated.

On the whole, all groups had similar distribution of amino acid contacts with Leucine – Leucine contact occurring with the highest probability. The majority of all Leucine contacts occurred with high probability giving rise to a dominant hydrophobic interaction in all groups. There were some contacts which occurred with high probability while being on different ends of the hydrophobicity scale. There were also proteins which deviated from the group and overall norm by a significant amount. These observations lead the way for future research.

5 FUTURE WORK

Future work with regard to this project includes the mapping of the calculated amino acid contact distribution back to protein sequence. Information on protein structure is not as widely available as that on protein sequence. Therefore calculation of the contact distribution with just information on the protein sequence would be useful. In addition, patterns in contact residues could be investigated. Given a contact residue pair any patterns with regard to its primary sequence could be investigated. For example, it could be investigated if the Leucine – Leucine contact is result of any specific primary sequence pattern.

REFERENCES

[1] Yanay Ofran and Burkhard Rost, *Analysing six types of Protein-Protein Interfaces*, J Mol Biol, 325, 377-387, (2003)
Available From:
http://cubic.bioc.columbia.edu/papers/2003_inter_ana/paper.html#ref1

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. *Nucleic Acids Research*, **28** pp. 235-242 (2000)
Available From:
http://www.rcsb.org/pdb/

[3] PFam: PFam Home Page
Available From:
http://www.sanger.ac.uk/Software/Pfam/