BioMetaNet: Meta-Network model for human lymphoblastoid cell lines representing complete biological interactome

Kaustav Sengupta[1,2], Michał Denkiewicz[1,3], Anup Kumar Halder[4], Subhadip Basu[4], Dariusz Plewczyński[1,3,5]

Email : k.sengupta@cent.uw.edu.pl, m.denkiewicz@cent.uw.edu.pl, dariuszplewczynski@cent.uw.edu.pl

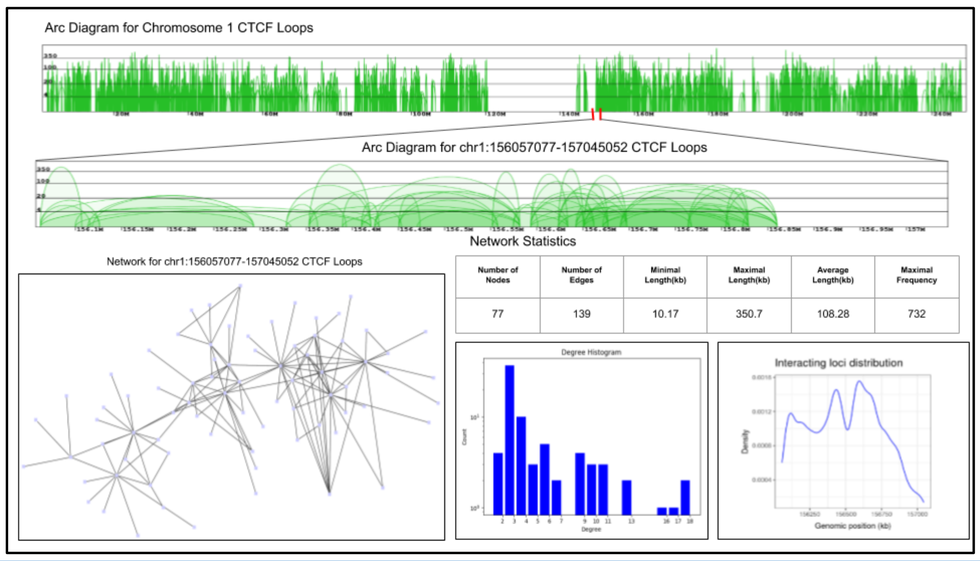
1. *Center of New Technologies, University of Warsaw, Poland*
2. *Faculty of Mathematics Informatics and Mechanics, University of Warsaw, Poland*
3. *Faculty of Mathematics and Information Science, Warsaw University of Technology, Poland*
4. *Department of Computer Science and Engineering, Jadavpur University, Kolkata, India*
5. *Computer Science Department, University of California, Davis, CA, United States*

Introduction

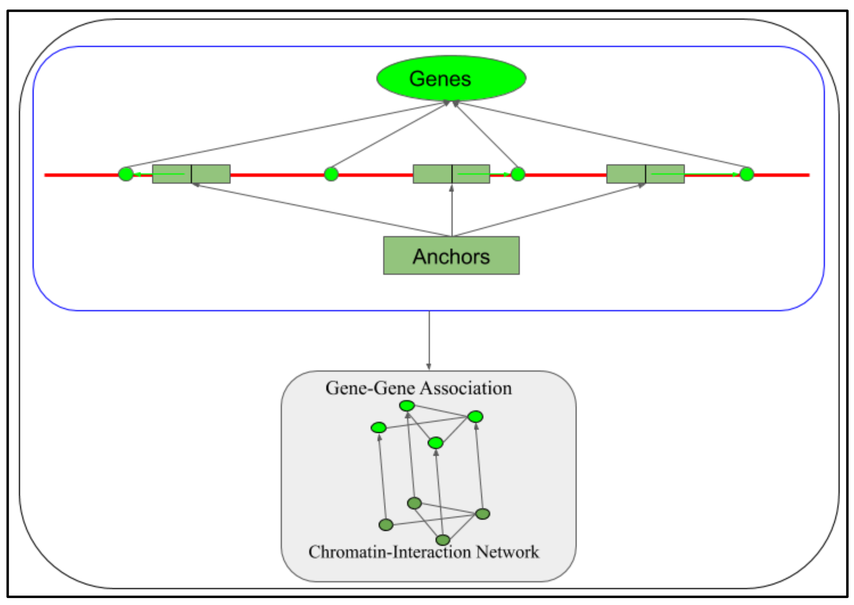
In recent years, with the development of high throughput methods, researchers obtained access to a vast array of biomolecular interaction data. Most of these biological data can be represented as networks or graphs. Thus, network analysis is becoming a powerful tool for modeling biological systems. We propose a meta-network representation of the complete map of DNA pairwise interactions for human lymphoblastoid cell lines combined with information about encoded proteins and metabolic pathways. In a single graph (meta-network) we integrate multiple biological networks, namely, Chromatin Interaction Network (CIN), Genomic Association Network (GAN), Protein-Protein Interaction Networks (PIN), Gene Ontology (GO) terms, and metabolic pathways. Thus cheating the meta-network connecting 3D chromatin interaction to functionality.

Methods

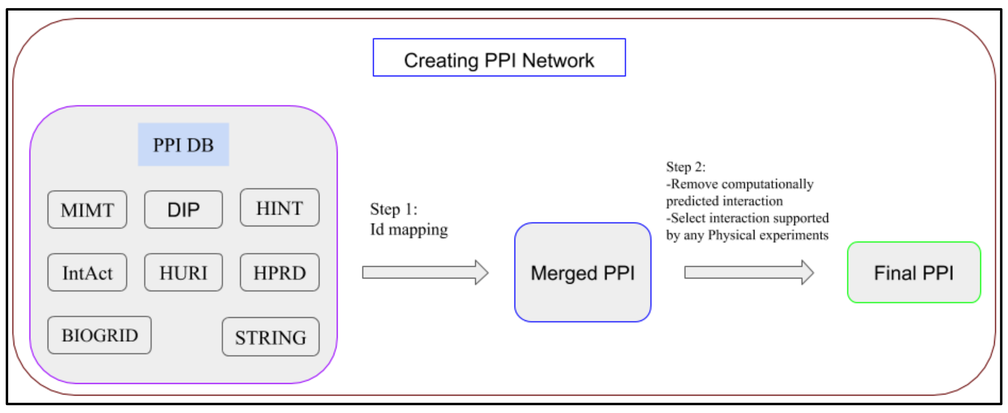
Chromatin Interaction Networks (CIN)

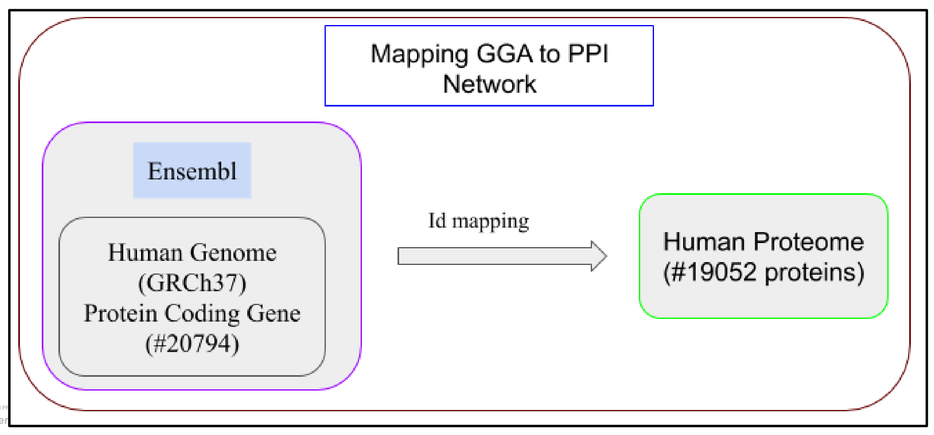


Gene-Gene Association Networks (GAN)

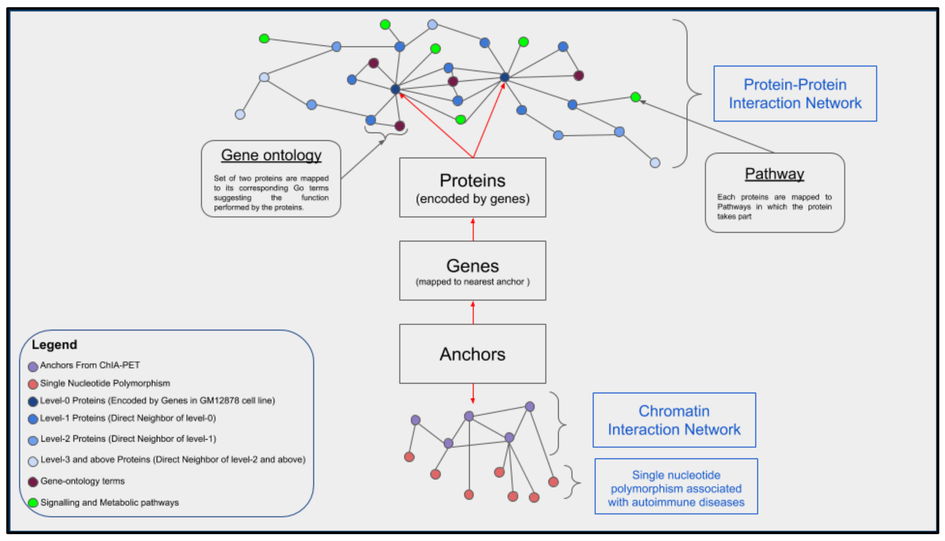


Protein-Protein Interaction Network (PIN)

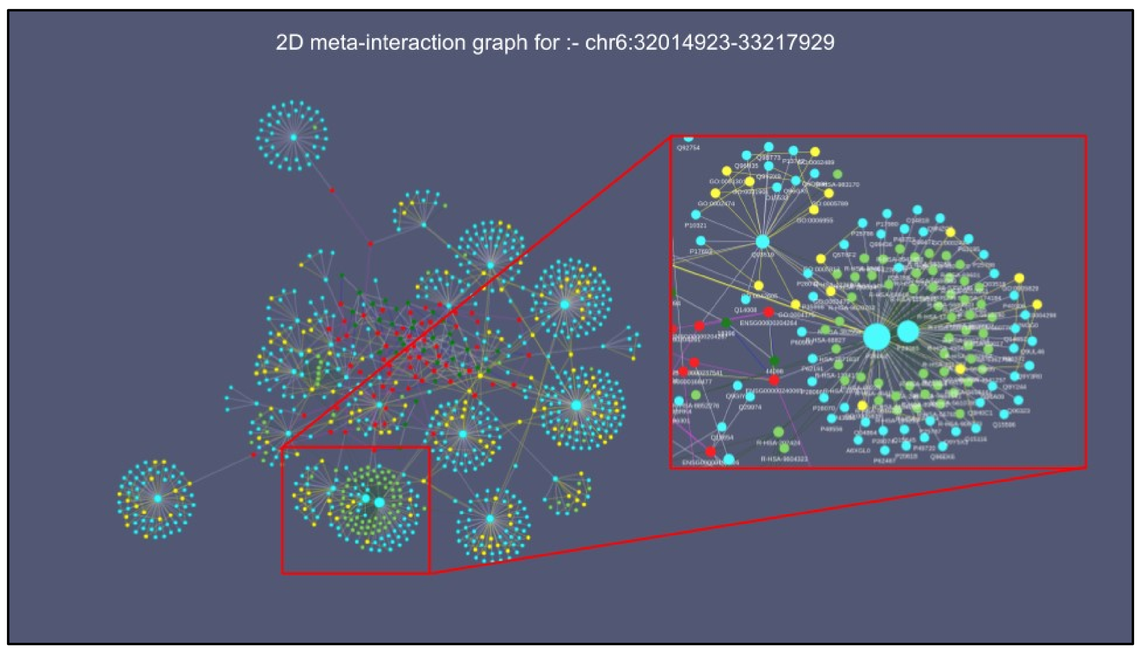




Gene-Ontology (GO), Pathways and Single Nucleotide Polymorphism (SNPs) Mapping



Results



We analyzed the meta-network and found proteins P28062 and P28065 encoded by genes PSMB8 and PSMB9, present in location chr6:32014923-33217929, share around 60 pathways which are higher than the average concentration of metabolic pathways shared between two proteins.

Critically, the genes PSMB8 and PSMB9 are also connected by proximity with HLA genes and TAP genes using the proteomic networks. The protein P28062 and P28065 are two of the 17 essential subunits (alpha subunits 1-7, constitutive beta subunits 1-7, and inducible subunits including beta1i, beta2i, beta5i) that contribute to the complete assembly of the 20S proteasome complex.

Conclusion

The meta-network can give us insigths into the interactions between genomic, proteomic and chromatin (structural) networks. In particular: the proteins P28062 and P28065, due to a large numer of shared pathways and the proximity of their encoding genes to the known autoimmune-related genes, can be critical for studies of autoimmune disease. Moreover, the presence of essential genes and proteins, the study of genome rearrangements in from of structural variants in this region can give us novel insights into the study of autoimmune diseases.

In conclusion, our meta-network model can be instrumental in getting a complete picture of biological functionality linked with 3D chromatin interactions. The network can also be extended to incorporate Structural Variants which can provide an idea of how functionality varies with the larger genome rearrangement.

Acknowledgement

This work has been supported by Polish National Science Centre (2019/35/O/ST6/02484), Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund (TEAM to DP). The work was co-supported by European Commission Horizon 2020 Marie Skłodowska-Curie ITN Enhpathy grant 'Molecular Basis of Human enhanceropathies'; and National Institute of Health USA 4DNucleome grant 1U54DK107967-01 "Nucleome Positioning System for Spatiotemporal Genome Organization and Regulation".

References

1.Anup Kumar Halder, Michał Denkiewicz, Kaustav Sengupta, Subhadip Basu, Dariusz Plewczynski,Aggregated network centrality shows non-random structure of genomic and proteomic networks,Methods,2019,ISSN 1046-2023,https://doi.org/10.1016/j.ymeth.2019.11.006.

2.Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, Michalski P, Piecuch E, Wang P, Wang D, Tian SZ, Penrad-Mobayed M, Sachs LM, Ruan X, Wei CL, Liu ET, Wilczynski GM, Plewczynski D, Li G, Ruan Y. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell. 2015 Dec 17;163(7):1611-27. doi: 10.1016/j.cell.2015.11.024. Epub 2015 Dec 10. PMID: 26686651; PMCID: PMC4734140.

3.Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyras, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H. R., Iyer, V., Jekosch, K., … Clamp, M. (2004). An overview of Ensembl. Genome research, 14(5), 925–928. https://doi.org/10.1101/gr.1860604 UniProt Consortium (2008). The universal protein resource (UniProt). Nucleic acids research, 36(Database issue), D190–D195. https://doi.org/10.1093/nar/gkm895

MASTERMIND: The Best Linear Model to Accurately Determine Monoisotopic Mass

Piotr Radziński, Michał Startek, Anna Gambin

Faculty of Mathematics, Informatics and Mechanics, University of Warsaw

Abstract

Nowadays, monoisotopic mass is used to be an important feature in top-down proteomics. Knowing the exact monoisotopic mass enables precise and quick protein identification in large protein databases. However, only in spectra of small molecules monoisotopic peak is visible, for bigger molecules position of the peak have to be predicted. By improving prediction of the peak, we contribute to more accurate identification of molecules, what is crucial in fields such as chemistry and medicine. In this work we present MASTERMIND algorithm, that is a two-step procedure to predict monoisotopic mass for proteins with 8-400 kDa mass range. The first step is to approximate monoisotopic mass by linear regression based on average mass and variance of a given spectrum. The second step rounds linear prediction to the closest point which is reliable to be a peak in the spectrum. For 96.6% of proteins, prediction error is below 0.2 ppm, what is approx. 30% better than in recently proposed MIND tool. Our algorithm was implemented in python, data analysis was performed in R. Proteins to learn the model comes from Uniprot database, their theoretical spectra were calculated by use of IsoSpec structure calculator.

MASTERMIND algorithm

I. Initial prediction

At the beginning, we calculate initial prediction of monoisotopic mass, by use of spectrum's average mass and variance:

$$\hat{M}_{\text{mono}} = \beta_0 + \beta_{\text{avg}} \cdot M_{\text{avg}} + \beta_{\text{var}} \cdot M_{\text{var}}.$$

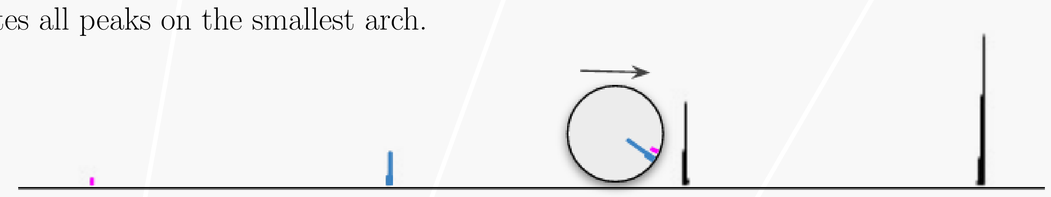
Prediction is not good enough for practical use, however, for 96.6% proteins prediction error is smaller than 0.5 Da, what is crucial for our algorithm. We want to round initial prediction to closest point on the grid

$$\mathcal{W}(\zeta, \Delta) = \{\zeta n + \Delta : n \in \mathbb{N}\},$$

which determine where peaks that are not visible on spectrum should be.

II. Estimation of the grid step ζ

Grid step ζ, is equivalent to circumference of circle, that rolled through spectrum concentrates all peaks on the smallest arch.



Mathematicaly, we have

$$\zeta_0 = \operatorname*{argmin}_{\zeta \in \mathbb{R}} \operatorname{Var} P_\zeta(\mathcal{S}),$$

where

$$P_\zeta(z) = \frac{\zeta}{2\pi i} \log\left[\exp\left(\frac{2\pi i z}{\zeta} - i\operatorname{Im}\left[\log\left(\sum_{p \in \mathcal{S}} p^{\text{prob}} \cdot \exp\left(2\pi i p^{\text{mass}}/\zeta\right)\right)\right]\right)\right].$$

To avoid long calculation for each protein, we trained linear model that gives ζ_0 based on protein average mass

$$\hat{\zeta} = \gamma_0 + \gamma_{\text{avg}} \cdot M_{\text{avg}}.$$

III. Estimation of the grid shift Δ

When we have $\hat{\zeta}$, we calculate grid shift, to fit the grid into spectrum

$$\hat{\Delta} = \operatorname*{argmin}_{\Delta \in [0, \hat{\zeta}]} \sum_{p \in \mathcal{S}} p^{\text{prob}} \cdot \min_{w \in \mathcal{W}(\hat{\zeta}, \Delta)} |p^{\text{mass}} - w| = \operatorname{Re}\left[\frac{\hat{\zeta}}{2\pi i} \log\left(\sum_{p \in \mathcal{S}} p^{\text{prob}} \cdot \exp\left(2\pi i p^{\text{mass}}/\hat{\zeta}\right)\right)\right].$$

IV. Final prediction

To obtain final prediction, we round initial prediction to closest point on the fitted grid, and apply slight correction

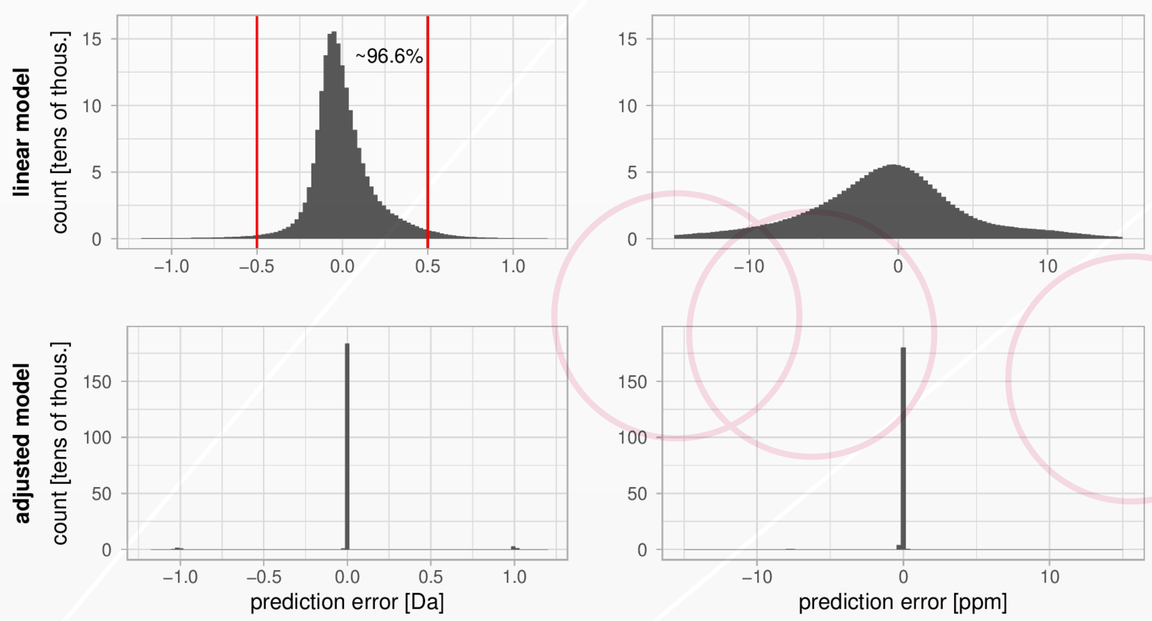
$$\hat{\hat{M}}_{\text{mono}} = \operatorname*{argmin}_{w \in \mathcal{W}(\hat{\zeta}, \hat{\Delta})} \left|w - \hat{M}_{\text{mono}}\right| + \lambda \cdot \hat{M}_{\text{mono}}.$$

Data & Tools

- Chemical formulas used to train models comes from **Uniprot** database;
- Their spectra were calculated by **IsoSpec** structure calculator;
- MASTERMIND algorithm was implemented in python, data analysis was performed in R. To calibrate linear models we used 10-fold cross-validation;

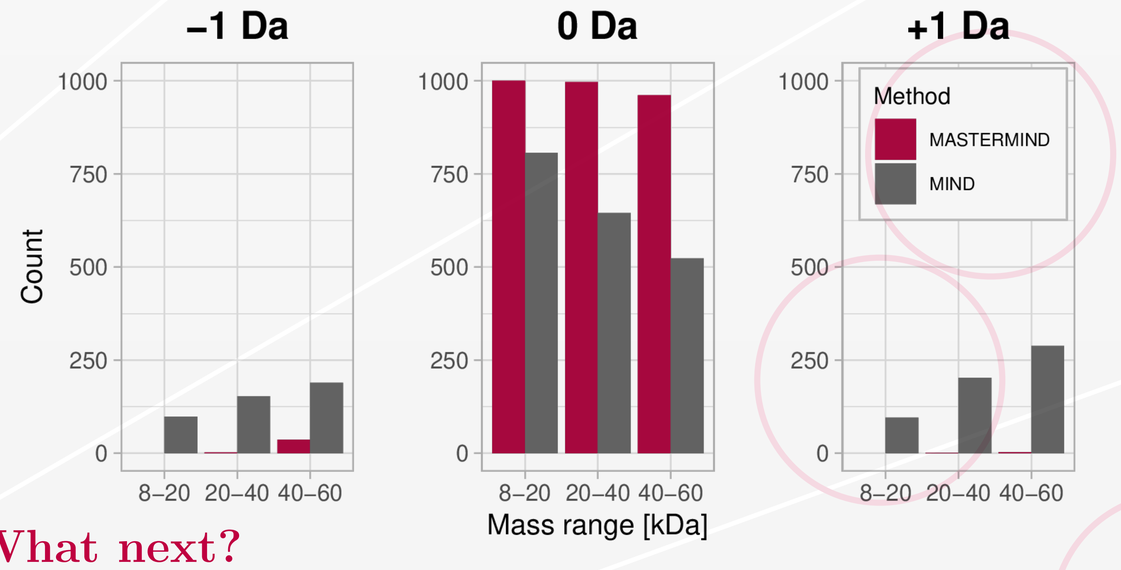
This research is supported by the Polish National Science Center grants 2018/29/B/ST6/00681 and 2017/26/D/ST6/00304.

How rounding improves prediction?



Comparison with MIND

- MIND prediction is based on the most-abundant peak, MASTERMIND is based on average peak and variance;
- MASTERMIND is close to true monoisotopic mass in **96.6%** versus 66.5% for MIND;
- MASTERMIND is better in every mass range it was compared with MIND, and is trained on bigger mass range;
- MASTERMIND loses accuracy fast, when spectrum resolution is getting worse;



What next?

- Elaborate a method, that finds average mass and variance regardless of spectrum resolution;
- Test MASTERMIND on real spectra;

References

Mateusz K. Łącki, Michał Startek, Dirk Valkenborg, Anna Gambin, 2017, *IsoSpec: Hyperfast Fine Structure Calculator*, Analytical Chemistry, vol. 89(6).

Frederik Lermyte *et al.*, 2019, *MIND: A Double-Linear Model To Accurately Determine Monoisotopic Precursor Mass in High-Resolution Top-Down Proteomics*, Analytical Chemistry, vol. 91(15).

Contact us!

pmradzinski@mimuw.edu.pl michal.startek@mimuw.edu.pl

A novel approach to search for interdigitated proteins - unusual domain swapped topology

Mateusz Skłodowski[1], Joanna M. Macnar[1,2], Dominik Gront[1]

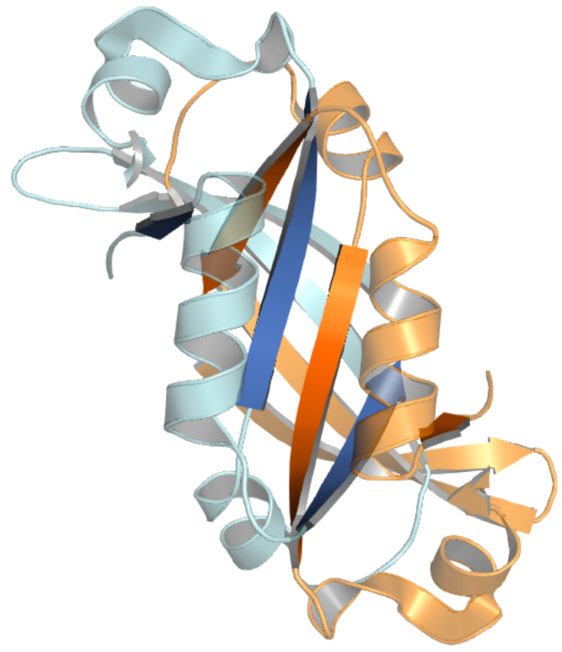
1 Faculty of Chemistry, University of Warsaw, Pasteura 1, Warsaw, Poland

2 College of Inter-faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Stefana Banacha 2C, Warsaw, Poland

Introduction

Interdigiteted motives are specific cases of protein domain swapping [1] including secondary structures from two different polypetide chains creating a single beta sheet. Additionally, interdigiteded stuctures consist of interchangeable occurrence of beta strands from different chains in beta-sheet.
In our work we search Protein Data Bank[2] for proteins that have the motive described earlier. For this task we used BioShell [3], [4] and graph theory. For further analysis, a group of proteins with the longest six-element beta sheet was adopted, in which their structural, sequential and functional similarity was studied.

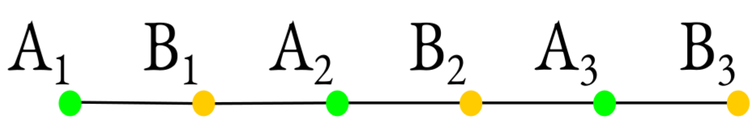
Interdigited protein example



Protein with six-element interdigitated beta sheet - AF2331[5]. *Darker colors represent secondary structures involve in motive.*

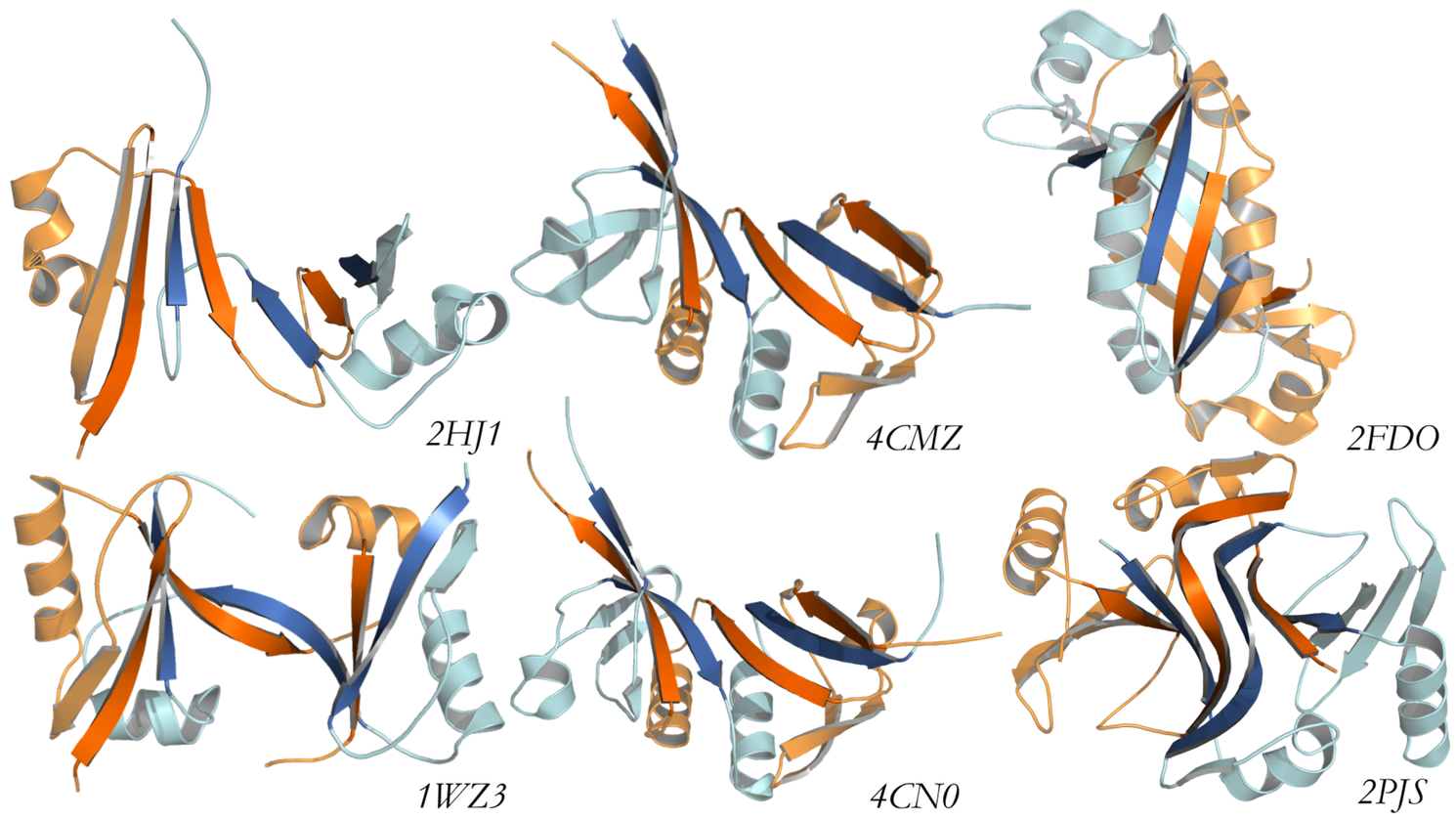
Graph theory application

In our project we applied graph theory to describe interactions between beta strands. For this work we state that each vertex of a graph is single beta strand. If the stands create a hydrogen bond, we assume an edge of the graph between them. To check if the beta sheet is interdigitated, we color the graph depending on the assignment of a beta strand to its protein chain. At this point, the depth-first search algorithm is used to gather information if interacting strands belong to different chains. The information collected also enables analysis in relation to the length of the motif.



Schematic application of the algorithm on the example of protein AF2331

Interdigited protein - examined group of proteins



Six proteins with six-element interdigitated beta sheet obtained by analysis

Conclusions

- Our approach has allowed us to identify new interdigitated proteins.
- We identify six proteins with six-element intedigiteted beta sheet.
- All of them are homodimers and their length does not extend beyond 120 aminoacids.
- We also identified a group of proteins with a smaller beta card. However, more research is needed in this subject.
- Another interesting topic is proteins, in which interdigitated beta sheets are formed by interactions of secondary elements from more than two chains.

Basic informations about examined group of proteins

Protein (PDB id.)	Year of publication	Sequence length [aa]	Homodimer?	Original organism	Crystal system	Resolution of measurement [Å]
1WZ3	2005	96	Yes	*Arabidopsis thaliana*	C2	1,8
2HJ1	2006	97	Yes	*Haemophilus influenzae*	C2	2,1
2PJS	2007	119	Yes	*Agrobacterium fabrum*	C2	1,85
4CN0	2014	97	Yes	*Homo sapiens*	C2	1,75
4CMZ	2014	92	Yes	*Homo sapiens*	C2	2,7
2FDO	2005	94	Yes	*Archaeoglobus fulgidus*	C2	2,4

References

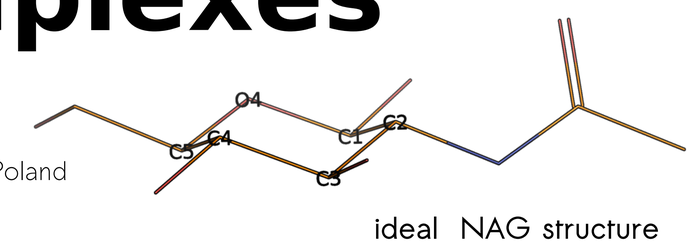
1 M. J. Bennett, S. Choe, and D. Eisenberg, “Refined structure of dimeric diphtheria toxin at 2.0 Å resolution,” Protein Sci., 1994, doi: 10.1002/pro.5560030911.
2 H. M. Berman et al., “The Protein Data Bank,” Nucleic Acids Research. 2000, doi: 10.1093/nar/28.1.235
3 D. Gront and A. Kolinski, “BioShell - A package of tools for structural biology computations,” Bioinformatics, 2006, doi: 10.1093/bioinformatics/btk037.
4 J. M. Macnar, N. A. Szulc, J. D. Kryś, A. E. Badaczewska-Dawid, and D. Gront, “Bioshell 3.0: Library for processing structural biology data,” Biomolecules, 2020, doi: 10.3390/biom10030461.
5 S. Wang et al., “The crystal structure of the AF2331 protein from Archaeoglobus fulgidus DSM 4304 forms an unusual interdigitated dimer with a new type of α + β fold,” Protein Sci., 2009, doi: 10.1002/pro.251.

BioShell software can efectivelly analyze rings in small compounds

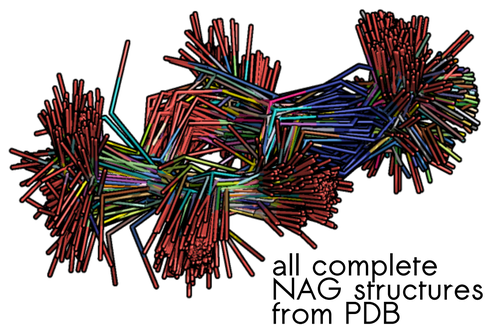
Analysis of small molecules parameters in ligand-protein complexes

Joanna M. Macnar [1, 2], *Wladek Minor* [3], *Dominik Gront* [1]

1. Faculty of Chemistry, Biological and Chemical Research Center, University of Warsaw, Warsaw, Poland
2. College of Inter Faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Poland
3. Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, USA



ideal NAG structure



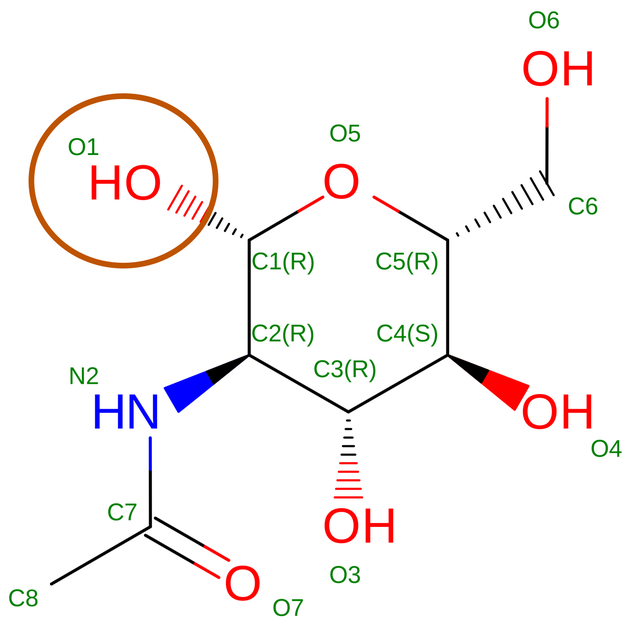
all complete NAG structures from PDB

Intro

Structural information about ligand-macromolecule complexes is critical for biomedical sciences. This analysis will lead to an improved library of restraint parameters and subsequently better refinement of ligand-protein complexes which contain 2-acetamido-2-deoxy-beta-D-glucopyranose (NAG).

Methods

We chose the most common small molecule from PDB which participates in a biological pathway and has one aliphatic ring. We found 5673 deposits and used BioShell package to analize their geometry.



Results

We analyzed 271 structures, thet were complete and determined by X-ray crystallography out of 5673 deposites that contained NAG ligands. A a referenceing X-ray crystallography. As a refence structure the `ideal.sdf` file form PDB was used.

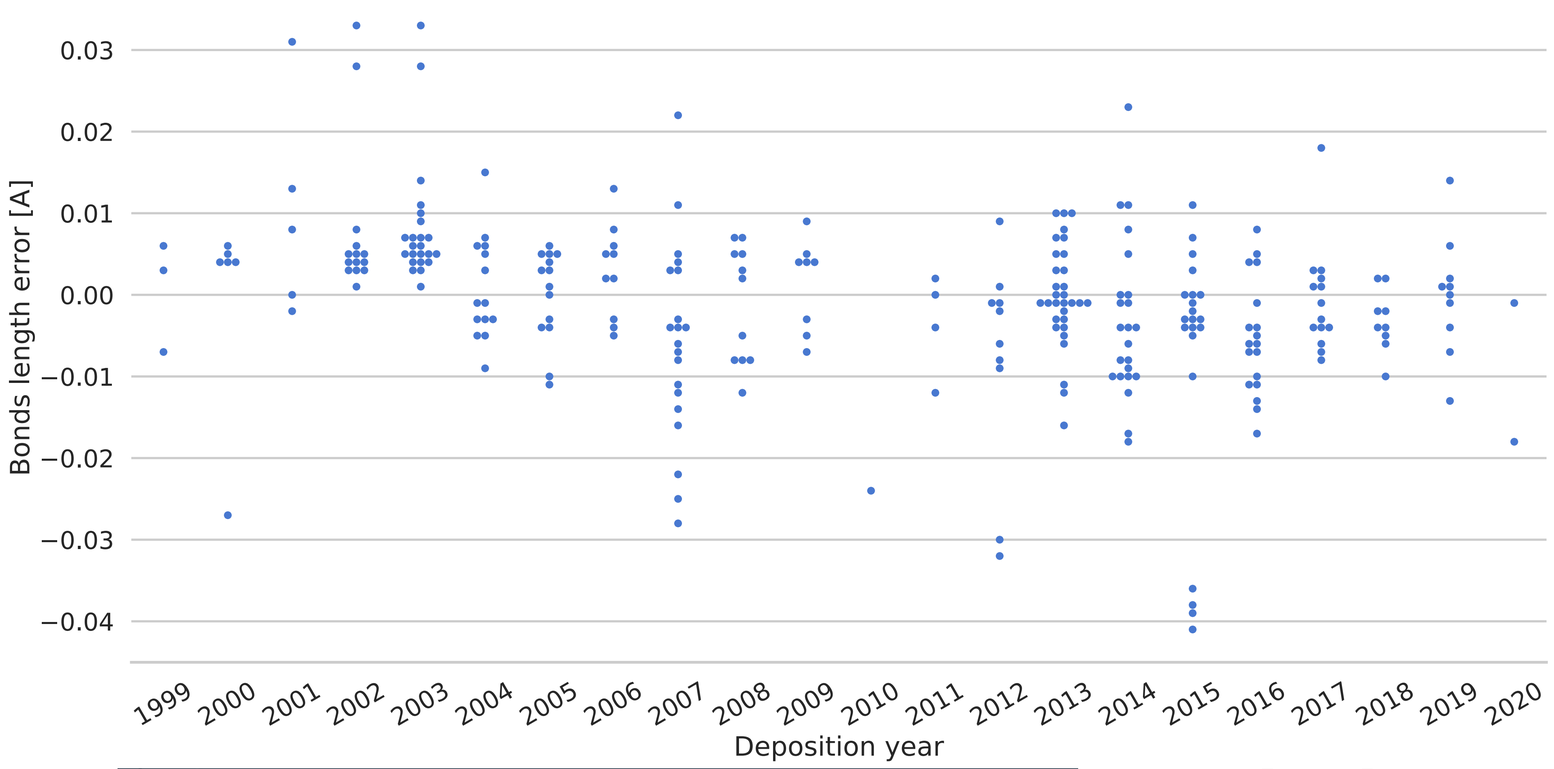


Figure 1

The average deviation of the bond lenght comparing to the ideal structure

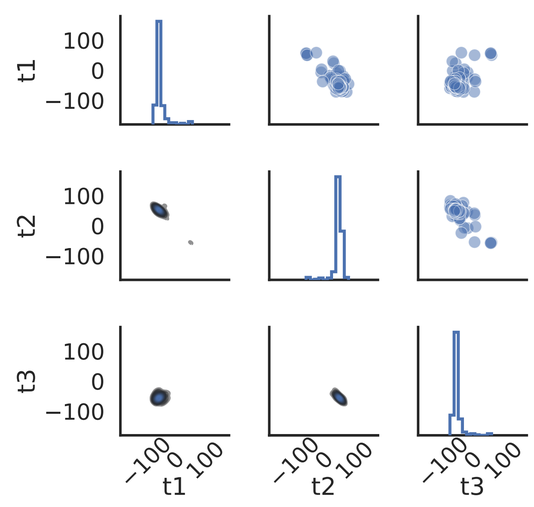


Figure 2

Scatter, KDE plots and histograms showing three subsequent torsion angles from a six member ring.

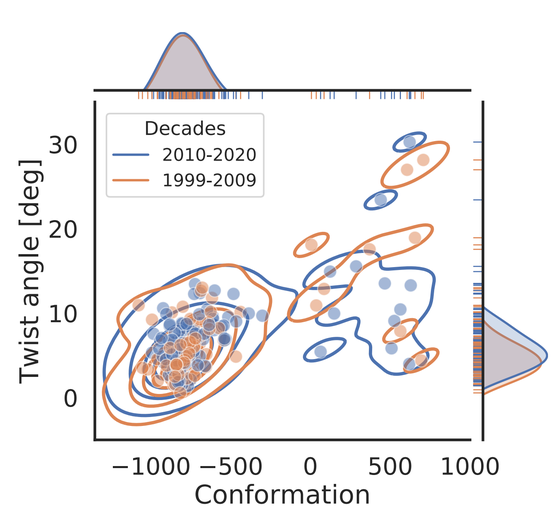


Figure 3

Conformational analysis of NAG rings showing improvementdeposit quality over time.

Conclusions

More reaserch is needed:

- The quality of NAG structures has remained roughly constant for 20 years
- Correlation to electron density map should be included for better analysis
- Missing ligand atoms are a common problem in deposits
- BioShell is a suitable package for ligand geometry analysis.

NATIONAL SCIENCE CENTRE POLAND

This research was funded by the National Science Centre (Poland) Grant No.2019/35/N/ST6/04459

UNIVERSITY OF WARSAW

Take a picture to **read more**