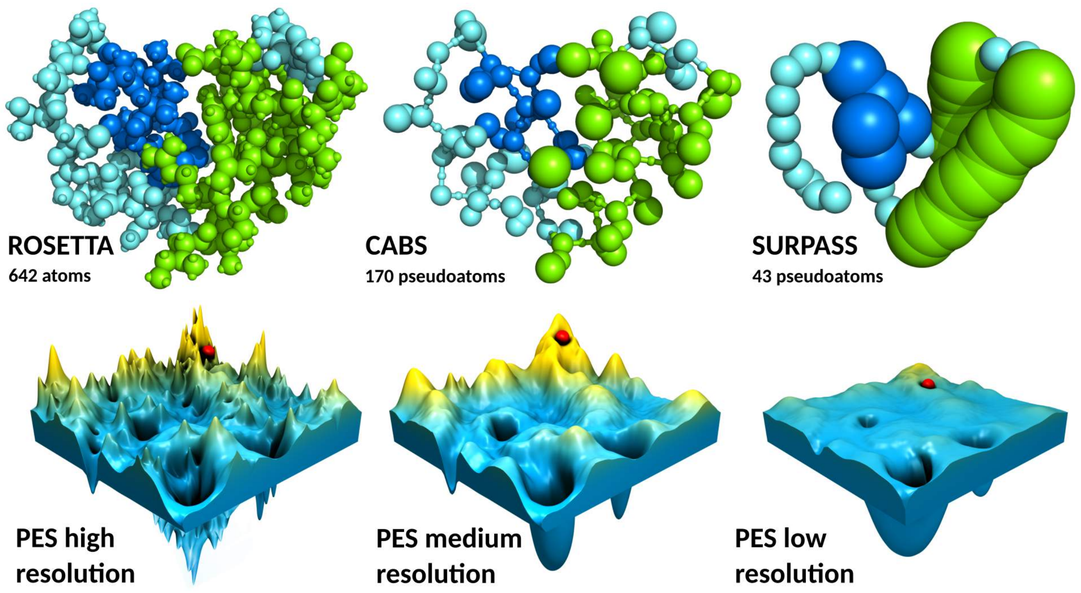
Multiscale Modeling of Protein Structure and Dynamics Using Coarse-Grained Models of Various Resolution

Aleksandra Elżbieta Badaczewska-Dawid

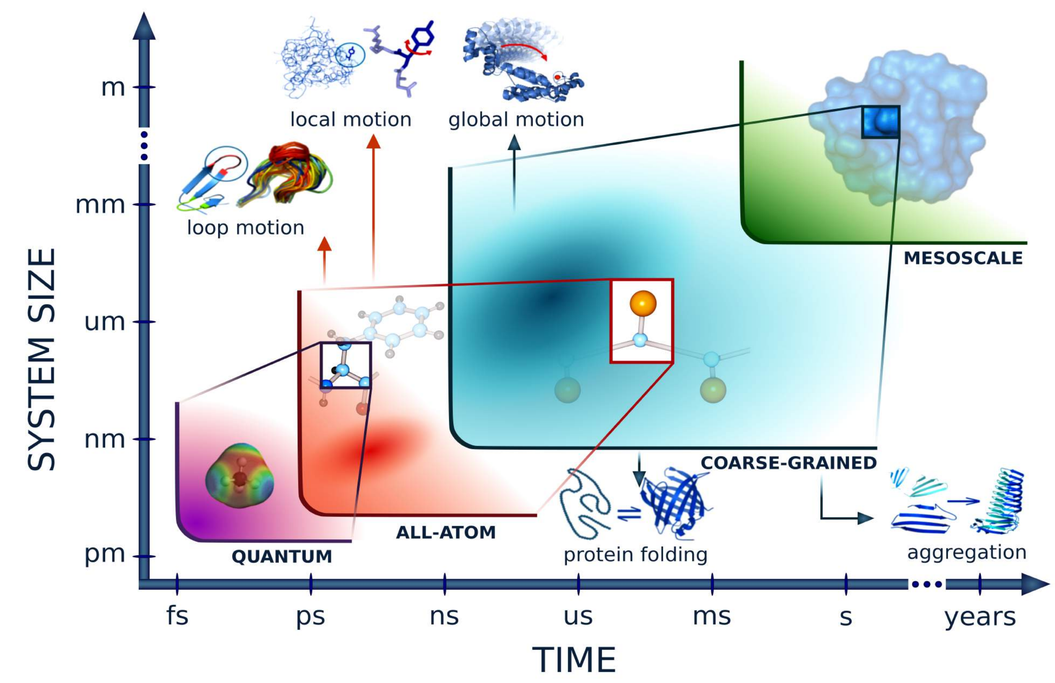
Department of Chemistry, Iowa State University, Ames, 50010 IA, U.S.; Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

COARSE-GRAINED PROTEIN MODELS



The coarse-grained models, their representation of protein chains, force fields, and sampling techniques must be carefully designed. In all coarse-grained models, the main purpose was to reduce the number of degrees of freedom. For this reason, pseudo atoms replace amino acid fragments or even entire amino acids. A broad spectrum of coarse-grained protein chain representations was proposed, starting with the simple lattice protein-like HP models or structurally more realistic low-resolution models like SICHO, by intermediate resolution coarse-grained models (e.g., CABS, UNRES) to almost exact coarse-grained protein models, like Rosetta or PRIMO. Medium-resolution CG models significantly expand the time scale and system size of molecular modeling. However, they struggle with de novo modeling of larger structures. Therefore, an efficient tool is needed to expand the range of de novo modeling of protein structure and dynamics by fast and efficient simulations of low-resolution structures.

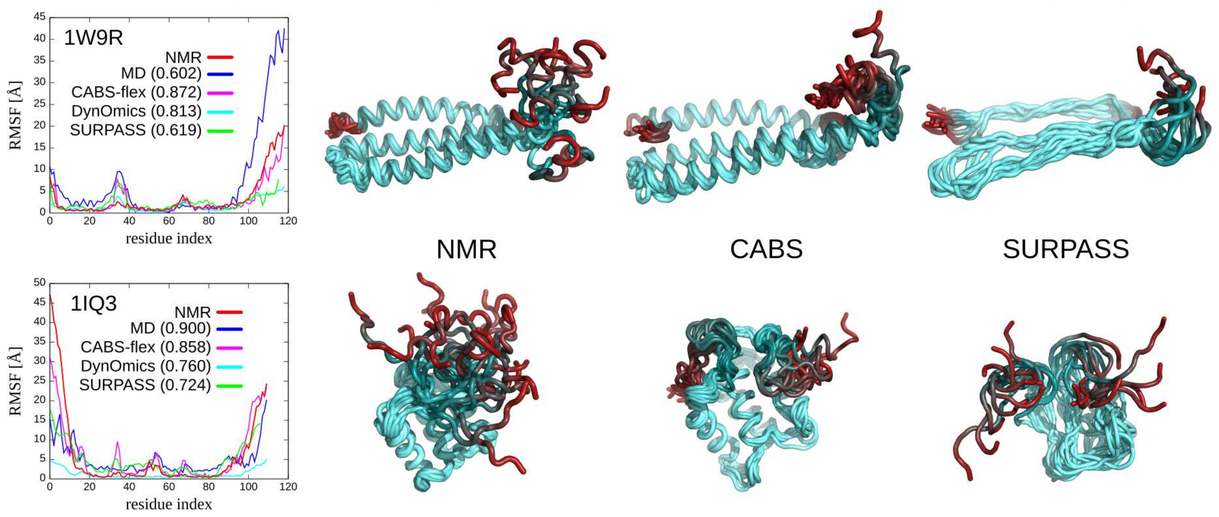
MULTISCALE MODELING



Classical atom-level molecular modeling can address many of the key tasks of structural biology, but its practical applications are still limited. This is a major reason why the development of coarse-grained protein modeling methods is needed. Coarse-grained models are computationally more effective and enable simulations of much longer time-scales and/or larger sizes of the systems studied. Multiscale methods that allow the transfer of information between various levels of granularity are more efficient and enable an analysis of larger systems on a longer time scale. Although successful multiscale modeling needs efficient and reliable algorithms for transferring information between calculations with different resolutions

CG MODELING OF PROTEIN DYNAMICS

Given the reports on the essential importance of protein dynamics for its biological function, we have studied the local flexibility of protein near the folded state. In the comprehensive study of 140 globular protein dynamics, we have applied various coarse-grained approaches: ENM-based modeling technique (DynOmics) and two representative simulation tools: medium-resolution CABS model and low-resolution SURPASS model. The proposed protocol succeeded in capturing the experimentally determined features (from NMR ensembles) of the investigated systems.

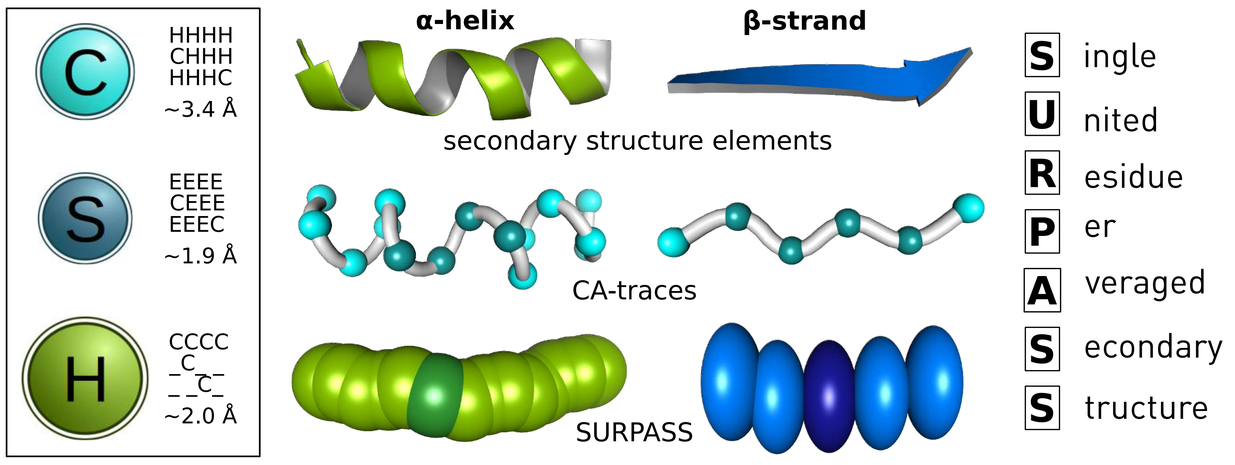


Due to its computational efficiency, SURPASS can be used for modeling long-time dynamics and large-scale structural transitions in protein systems that are significantly bigger than those tractable by the coarse-grained modeling tools of higher resolution. The models such as SURPASS can be useful as part of multiscale molecular modeling schemes. In such a scheme, SURPASS simulations can provide a collection of protein-like low-resolution starting structures, and these could be used for more accurate methods, e.g., as an input to replica-exchange simulations with a medium-resolution CG model (for example, CABS). Intermediate resolution structures can be finally subjected to all-atom reconstruction and MD refinement/scoring simulations.

SURPASS MODEL

LABORATORY of THEORY of BIOPOLYMERS http://biocomp.chem.uw.edu.pl/tools/surpass Andrzej Kolinski Research Group

RESEARCH PEOPLE PUBLICATIONS TOOLS CONTACT

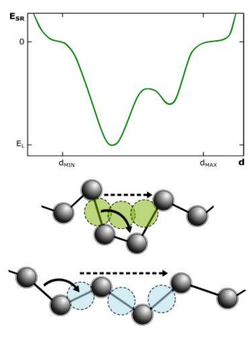


SURPASS is a low-resolution, deeply coarse-grained model of protein structure. The number of pseudo residues representing protein structure corresponds to the length of the protein sequence. The main idea behind the model is based on a unique generalization of the local geometry of a polypeptide chain. Namely, positions of pseudo atoms are defined by averaging the coordinates of the four consecutive α-carbons along the chain. These four-residue fragments are replaced by a single center of interactions. The choice of four-residue averaging is crucial for the geometry of the model. In contrast to other short fragments of different lengths, only the four-residue averaging leads to an almost linear shape of the SURPASS fragments representing helices or ß-strands. This feature of the model results in simple and effective sampling schemes. The SURPASS representation assumes three types of pseudo atoms depending on secondary structure assignment: H (helical), S (ß-strand), C (coil-like).

SURPASS FORCE FIELD

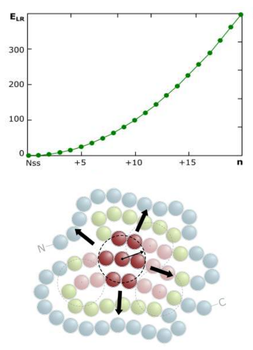
Designing and derivation the force field for a coarse-grained model is always a key point for its performance. A combination of the statistical potentials defines the knowledge-based SURPASS force field. They describe local structural regularities characteristic for most globular proteins. The generic terms are basically sequence-independent and are encoded non-directly via secondary structure assignment. The solvent is treated implicitly, and its effects (water with other small molecules or a membrane environment for transmembrane proteins) are included directly in the statistical potentials that describe interactions between the united residues.
The specific interaction model distinguishes the protein-like SURPASS chain from a random polymer.

SHORT RANGE R12, R13, R14, R15



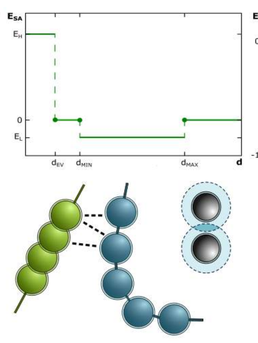
Distances and angles between atoms close along the sequence in polypeptide chains are highly restricted due to various short-range interactions, which provide the correct local geometry of the structure.

LOCAL REPULSIVE



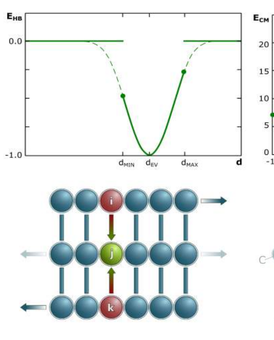
To prevent the excessive and non-physical collapse of a structure, generic local repulsions are needed. Using deeply buried elements derived from PDB, we estimated the number of neighboring SURPASS atoms.

SOFT-CORE ATTRACTION



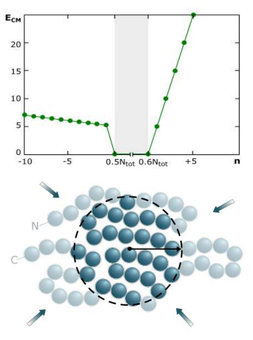
To avoid steric clashes between pseudo atoms distant in sequence but close in space, the excluded volume cut-off was derived. Contacts are rewarded only for specific distances between atoms of a given II-structure type.

HYDROGEN BOND



H-bonds between residues close to each other along the chain are treated implicitly. H-bonds between residues that are distant in the sequence (in the extended fragments) are modeled more directly.

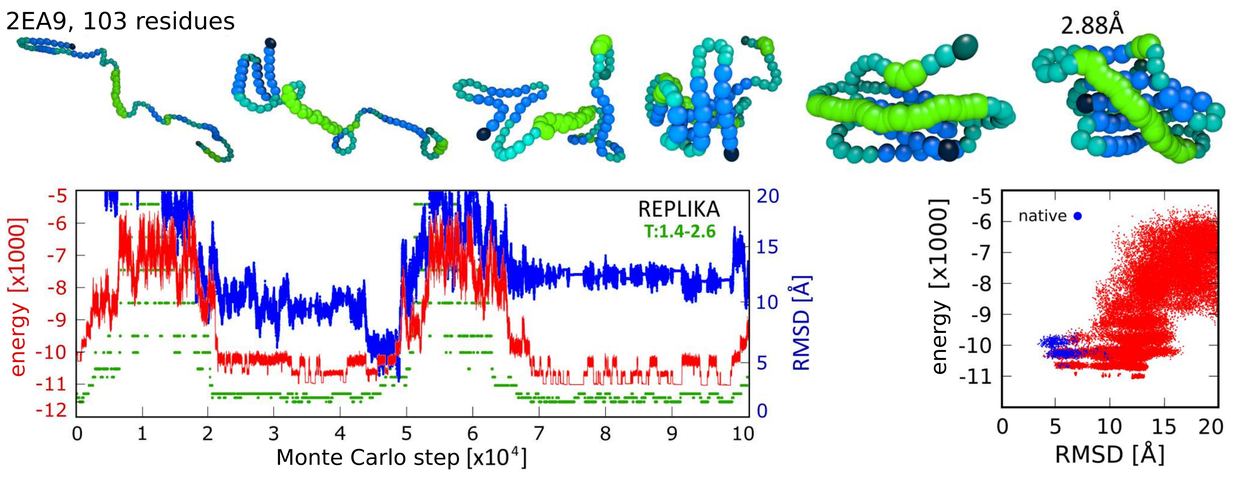
CENTRO-SYMMETRIC



To force the SURPASS chain to fold into globular topology, we used a simple centrosymmetric potential. Its purpose is to maintain a sufficiently high degree of packing of pseudo residues in the protein core.

SURPASS MODELING OF PROTEIN STRUCTURE

SURPASS model was used for replica-exchange Monte Carlo dynamics simulation of proteins, with secondary structure as the only sequence-dependent input data for the interaction model. The studied cases were a representative set of single-domain globular proteins. The set contained 7 helical proteins, 9 mostly ß-sheet, and 8 mixed alpha/beta proteins. In the test simulations presented here, the secondary structure assignments required by the model were taken directly from the PDB database. Replica exchange Monte Carlo simulations were performed with 12 replicas for each tested protein. The starting structures of all replicas had fully expanded the conformation of model chains.



The method efficiently samples the entire conformational space of polypeptide chains. Despite its deep simplification, the SURPASS model reproduces reasonably well the basic structural properties of proteins. Also, the accuracy of the resulting native-like models, measured by the RMSD between the generated chains and the SURPASS representation of experimental structures, is surprisingly good for such a level of coarse-graining. We demonstrated that different assignments and/or predictions of secondary structures are sufficient for enforcing cooperative formation of native-like folds of SURPASS chains for the majority of single-domain globular proteins. Simulations of globular protein structure assembly have shown that the accuracy of secondary structure data is usually not crucial for model performance.

S. Kmiecik, et.al., Chem. Rev. 2016, 116, 14, 7898–7936
A.E. Dawid, et.al., J. Chem. Theory Comput., 2017, 13(11), 5766–5779
A.E. Dawid, et.al., J. Chem. Theory Comput., 2018, 14(4), 2277–2287
S. Kmiecik, et.al., Int. J. Mol. Sci. 2018, 19(11), 3496
A.E. Badaczewska-Dawid, et.al., Methods Mol. Biol., 2020, vol 2165

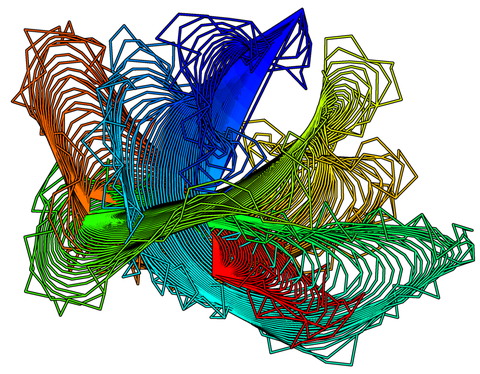
UNIVERSITY OF WARSAW

IOWA STATE UNIVERSITY

Algorithms and models for protein structure analysis

Aleksandra I Jarmolińska[1,2], Joanna I Sulkowska[1], Anna Gambin[2]

1Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097, Warsaw, Poland
2Faculty of Mathematics, Informatics, and Mechanics, Uniwersity of Warsaw , 02-097 Warsaw, Poland



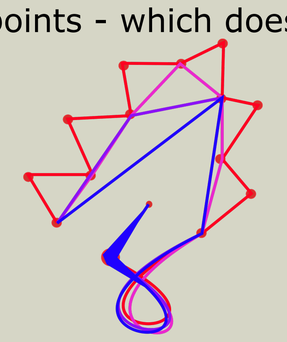
The last decade has seen a large increase in the number of studies related to protein topology. Currently, there are over 1500 known knotted or slipknotted protein chains and almost 10 000 protein links. Screening of available RNA structures has also found entanglements. Recent advances in the study of chromatin structure gave rise to new 3D models—many of which contain entanglements, including composite knots. Still, the subject of molecular entanglements remains relatively unknown to a lot of researchers, including those studying protein structures. One obvious reason is the steep learning curve for actually seeing the knots in a 3D structure visualization. Knot_pull (Jarmolinska et al, 2019) allows an easy analysis of topological intricacies by providing the user with a trajectory of smoothing steps—from the full structure, to the minimal number of coordinates preserving the original topology (with regard to fixed position of chain termini) — and the knot type (including separation of composite knots, and indication of any linking present) - without using the prevalent probabilistic approach.

Studying the sequences of entangled proteins also encounters problems - finding the most closely related protein family may require detecting the similarity based on sequence profiles, which are not easily (multiple-)aligned. To overcome this obstacle, I introduce two new heuristic for creating a multiple profile alignment, by using a modified Dijkstra's shortest path tree algorithm to find the maximum weight trace (Kececioglu, 1993) of a set of pairwise alignments. This allows for an easy, large scale comparison of loosely related protein groups.

Simplify

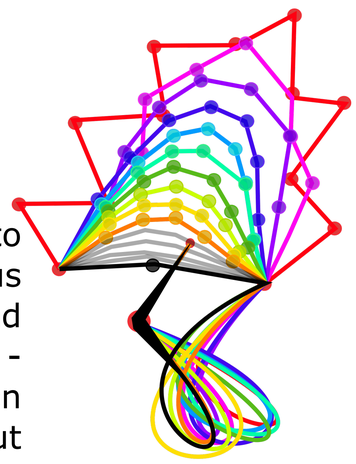
Previous approach

Structures are simplified by reducing the number of points - which doesn't necessarily extrude the ends.

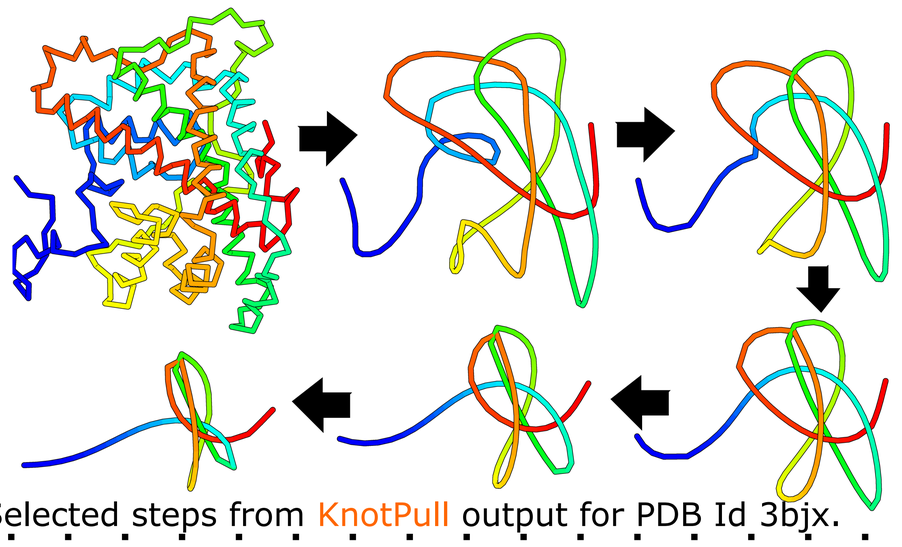


KMT

KnotPull allows all connections to be split or shortened thus tightening the entanglement - and making the ends stand out more - so that there exists a projection which could be closed without adding to the entanglement.

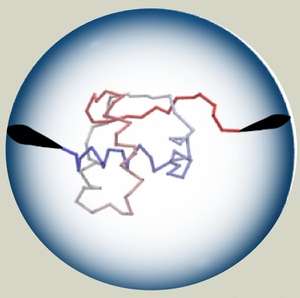


knot_pull



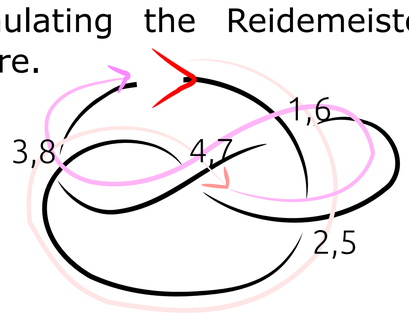
Selected steps from KnotPull output for PDB Id 3bjx.

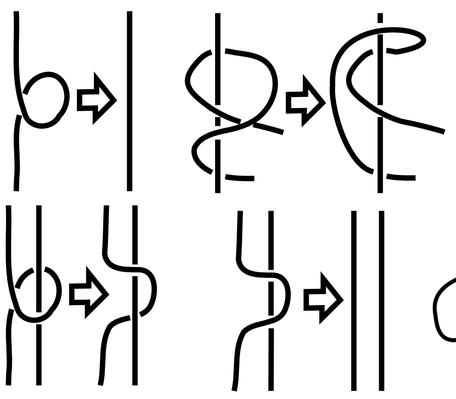
Recognize



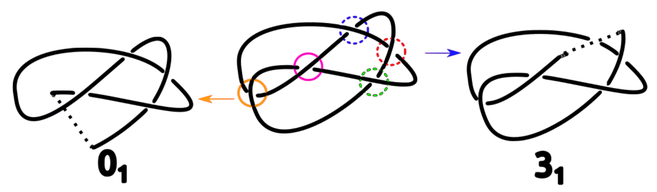
Knot type is calculated based on a polynomial, which requires a closed 3d curve projected on a plane. Open chain needs to be closed - randomly so as not to affect the results.

KnotPull uses the Dowker-Thistlethwaite code, which reads the structure as implicitly closed. This code is then simplified by calculations on the code itself simulating the Reidemeister moves on the structure.





Additinally, DT code recognizes different realizations (based on the location of the break - the temini) of the same knot type.



HHaligner

aligns multiple sequence profiles

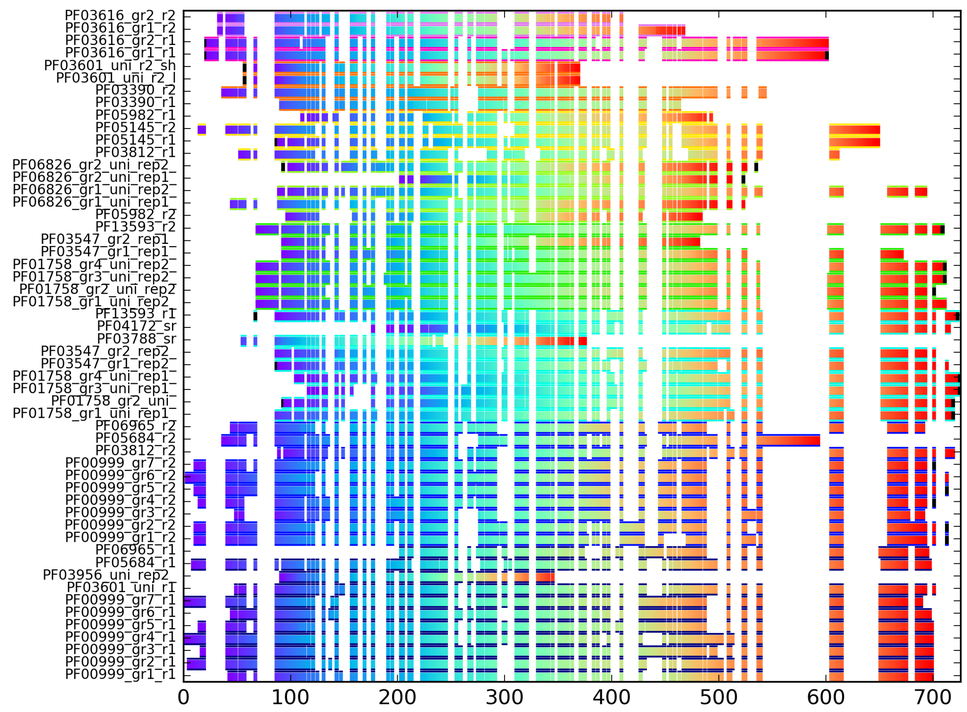
https://github.com/dzarmola/HHsearch-resuits-aligner

KnotPull

finds entanglements in 3D structures

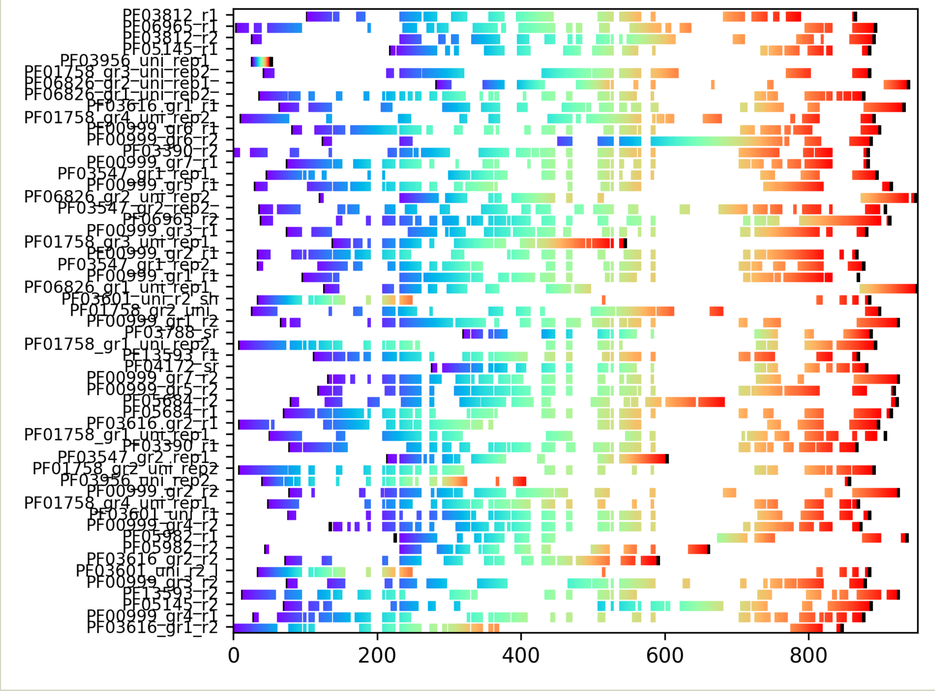
https://github.com/dzarmola/knot_pull

Multiple sequence alignment is a great way of analysing the similarities between related sequences - aligning their matching regions highlights the conserved characteristics. However, a single sequence is not enough to convey the diversity of a given protein - thus MSA cannot fully show the similarities between the entire families, like a multiple **profile** alignment could.



Right: MSA of the representatives of multiple distantly related families of transmembrane proteins.

Left: MPA of the profiles of the same



HHaligner creates a graph of positions in profiles connected based on their similiarity in pairwise profile alignments. Then, by heuristically optimizing the maximum weight trace, we can resolve the alignment.
Since the idea is to find the similarities, this approach additionally clears out any insertions specific to just one of the profiles.

Bibliography

Jarmolinska AI, Sulkowska JI, Gambin A, Bioinformatics (2020) - Knot_pull - python package for biopolymer smoothing and knot detection.

Kececioglu J, Annual Symposium on Combinatorial Pattern Matching (1993) - The maximum weight trace problem in multiple sequence alignment

Jarmolinska AI PhD thesis (2019) - Algorithms and models for protein structure analysis

CeNT CENTRE OF NEW TECHNOLOGIES

UW MM Wydział Matematyki, Informatyki i Mechaniki

INTEGRATIVE DATA ANALYSIS METHODS IN MULTI-OMICS MOLECULAR BIOLOGY STUDIES FOR DISEASE OF AFFLUENCE BIOMARKER RESEARCH

Anna Papież
Supervisor: Joanna Polańska
Data Mining Group, Institute of Automatic Control, Silesian University of Technology

Silesian University of Technology

Public Health England

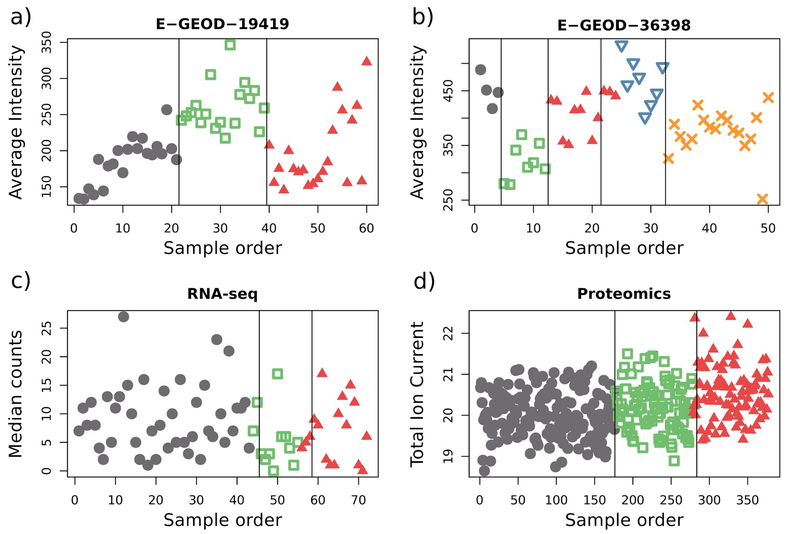
HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

AIM OF THE STUDY

The goal of this work was to investigate diverse approaches for high-throughput molecular biology integrative data analysis to enable the discovery of disease of affluence biomarkers. The research methodology comprises a thorough overview of existing approaches for data combination, merging, comparison, and joint analysis, as well as the development of new methods for handling multi-omics studies. The expected outcomes of this work include the establishment of novel tools and procedures tailored to the tasks of multi-platform and multi-omics data and result integration.

INTRA-EXPERIMENT INTEGRATION

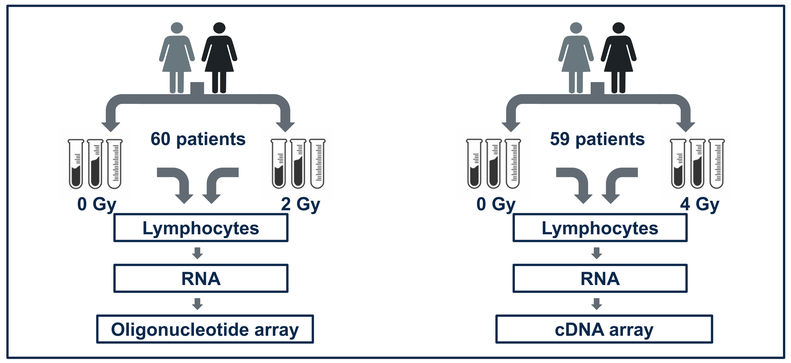
An original batch effect identification algorithm based on dynamic programming was proposed, as correcting for these effects constitutes a part of the intra-experiment data integration pipeline. The BatchI algorithm is based on partitioning a series of high-throughput experiment samples into sub-series corresponding to estimated batches. The dynamic programming method is used for splitting data with maximal dispersion between batches, while maintaining minimal within batch dispersion. The procedure has been tested on a number of available datasets with and without prior information about batch partitioning. Datasets with a priori identified batches have been split accordingly. Batch effect correction is justified by higher intra-group correlation. In the blank datasets, identified batch divisions lead to improvement of parameters and quality of biological information, shown by literature study and Information Content.



The BatchI algorithm's performance on identifying batch structure is proven to be highly efficient, and moreover, batch effect preprocessing entails potential new knowledge discovery in studied diseases and conditions. It is available to the scientific community as an R package.

INTER-PLATFORM INTEGRATION

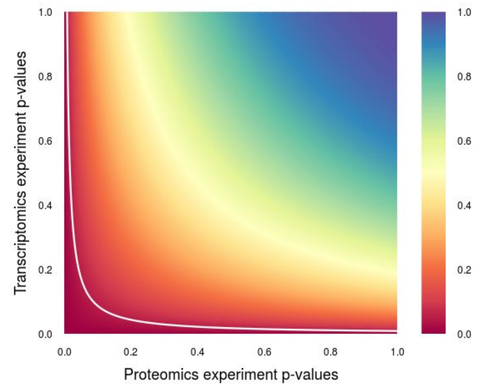
The analyzed data consist of two gene expression sets obtained in studies of radiosensitive and radioresistant breast cancer patients undergoing radiotherapy. The data sets were similar in principle; however, the treatment dose differed. It is shown that introducing mathematical adjustments in data preprocessing, differentiation and trend testing, and classification, coupled with current biological knowledge, allows efficient data analysis and obtaining accurate results. The tools used to customize the analysis workflow were batch effect filtration with empirical Bayes models, identifying gene trends through the Jonckheere–Terpstra test and linear interpolation adjustment according to specific gene profiles for multiple random validation.



The application of non-standard techniques enabled successful sample classification at the rate of 93.5% and the identification of potential biomarkers of radiation response in breast cancer, which were confirmed with an independent Monte Carlo feature selection approach and by literature references. This study shows that using customized analysis workflows is a necessary step towards novel discoveries in complex fields such as personalized individual therapy.

INTER-OMICS INTEGRATION

The goal of this part was to elucidate molecular mechanisms of radiation-induced IHD by integrating proteomics data with a transcriptomics study on post mortem cardiac left ventricle samples from Mayak workers categorized in four radiation dose groups (0 Gy, < 100 mGy, 100–500 mGy, > 500 mGy). The proteomics data originated from a label-free analysis of cardiac samples. The transcriptomics analysis was performed on a subset of these samples. Stepwise linear regression analyses were used to correct the age-dependent changes in protein expression, enabling the separation of proteins, the expression of which was dependent only on the radiation dose, age or both of these factors. Importantly, the majority of the proteins showed only dose-dependent expression changes. Hierarchical clustering of the proteome and transcriptome profiles confirmed the separation of control and high-dose samples. Restrictive (separate p-values) and integrative (combined p-value) approaches were used to investigate the enrichment of biological pathways.



Custom statistical integrative methods applied to a transcriptomics and proteomics data set on ischemic heart disease plutonium mine workers enabled discrimination of dose dependent protein expression changes from the age dependent changes and validation of pathways identified previously in the proteomic data.

This work was funded partially by:
- "DoktoRIS - Scholarship Program for Innovative Silesia", co-financed by the European Union under the European Social Fund
- POIG.02.03.01-24-099/13 grant funding and infrastructure:"GeCONiI - Upper Silesian Center for Computational Science and Engineering"
- National Science Centre Harmonia grant number DEC-2013/08/M/ST6/00924: "BioRadInt"
- National Science Centre OPUS grant number UMO-2015/19/B/ST6/01736: "BiTIMS"
- National Center for Research and Development PBS grantnumber PBS3/A7/29/2015/ID-247184 "MOLTEST BIS"