Machine Learning Characterization and Prediction of Intrinsically Disordered Proteins Interactions:
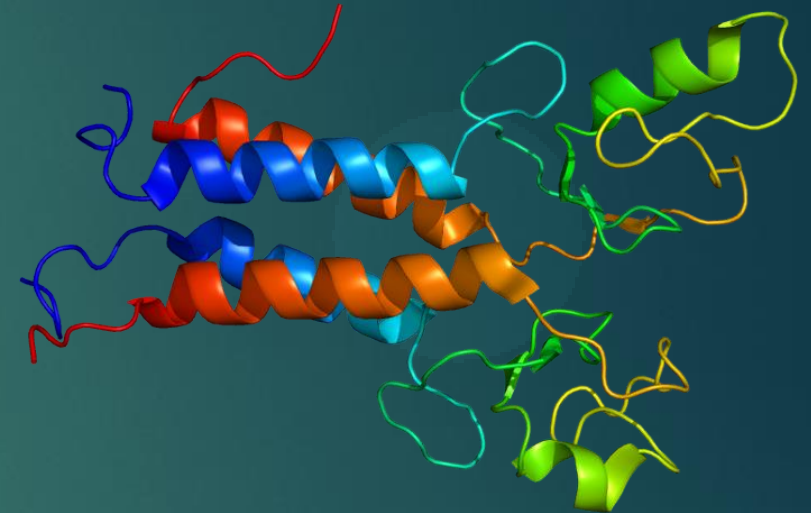
A Focus on BRCA1

ASHAY ATHALYE

MENTOR: DR. GIL ALTEROVITZ

5TH ANNUAL PRIMES CONFERENCE

MAY 17TH, 2015

BRCA1 (Breast Cancer 1, early onset)

- Tumor suppressor gene (Caretaker gene)
- BRCA1 protein: DNA repair and transcriptional regulation
- Mutation → malfunctioning protein
  - Cancer Development
 - Responsible for > 80% breast and ovarian cancers

Figure 1: PDB rendering based of model 1jm7

Clinical Significance: BRCA Mutations and Cancer

- BRCA mutation in 90% of familial breast and ovarian cancers
- BRCA mutation → **5x** breast cancer risk, **30x** ovarian cancer risk

Figure 2: Proportion of hereditary breast cancer cases caused by each gene

- BRCA1 mutation
- BRCA2 mutation
- All other known genes
- Unknown genes or multiple genes
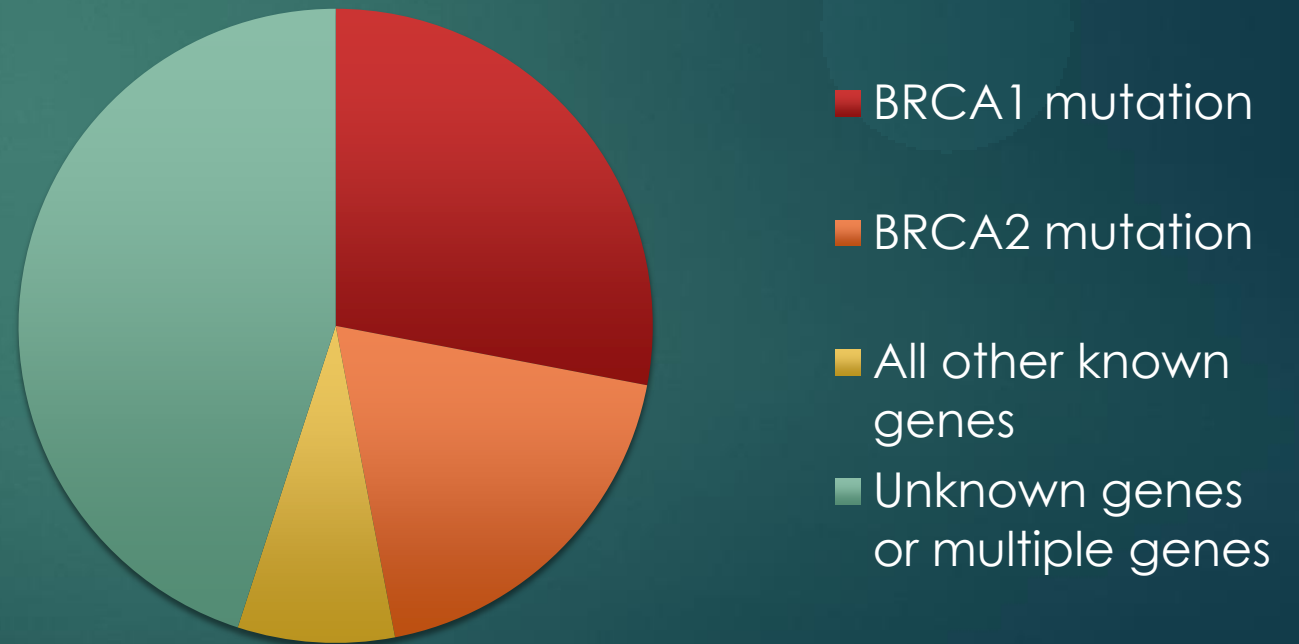
Problem

- Mutations in the BRCA1 lead to cancer

- Need to create a drug to replicate the behavior of the BRCA1 protein to ensure damaged DNA repair
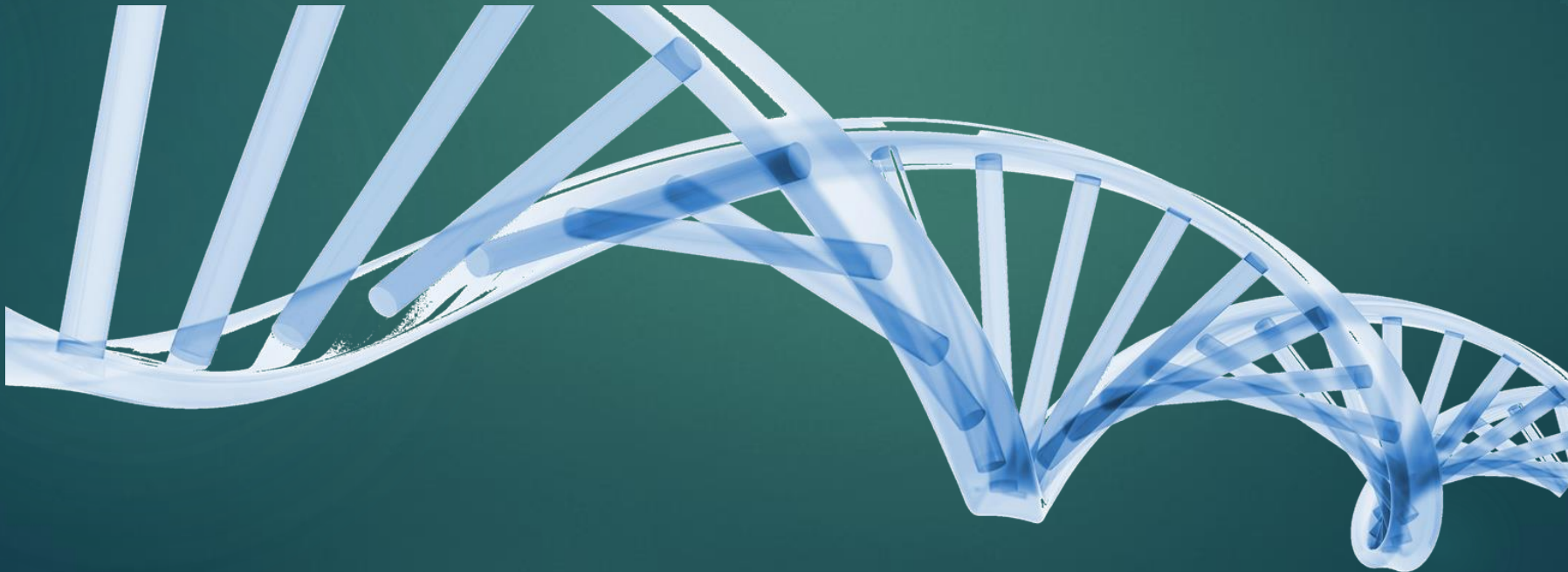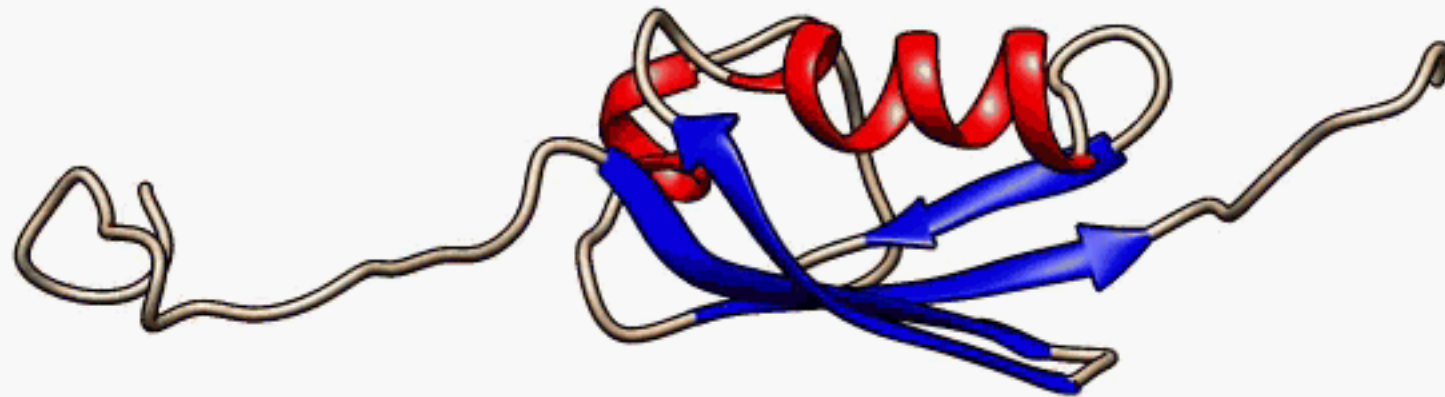
Figure 3: Conformational flexibility in PDB model 1a5r

(Majorek K, Kozlowski L, Jakalski M, Bujnicki, JM (December 18, 2008). *Prediction of Protein Structures, Functions, and Interactions.* John Wiley & Sons, Ltd.)

BRCA1 and Intrinsic Disorder

- 1863 amino acid protein
  - N-terminal RING domain (1 – 103; ~5%)
  - Central disordered region (104 –1645; ~79%)
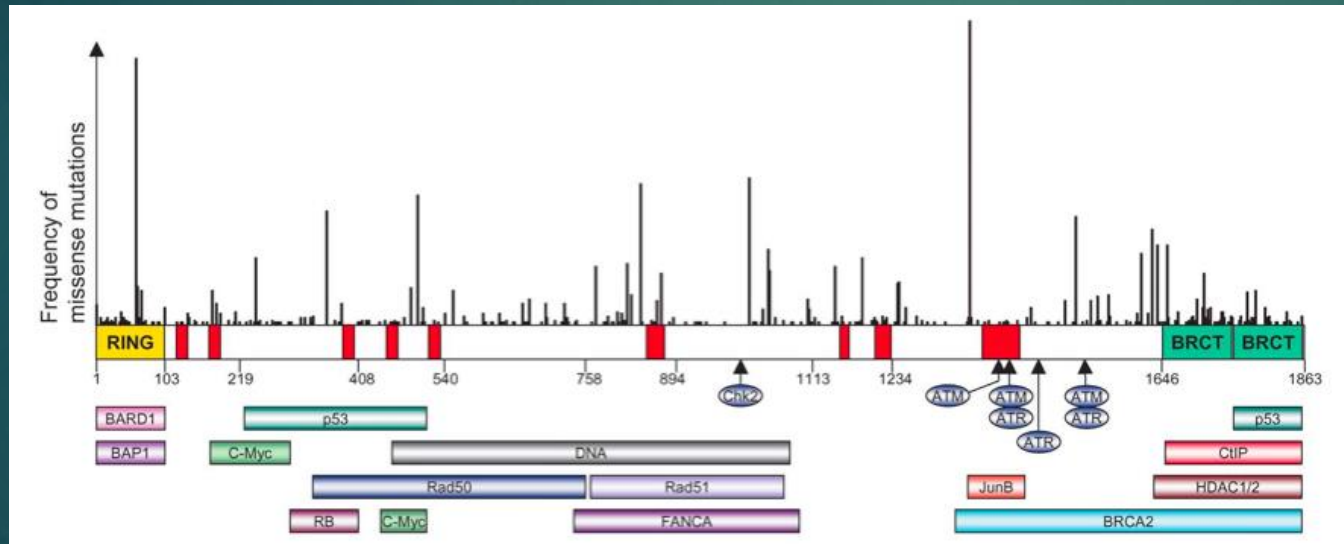 - C-terminal BRCT domains (1646 – 1863; ~11%)

Figure 4: Representation of BRCA1 protein with interactors

(Mark, W., et al (October 14, 2004). Characterization of Segments from the Central Region of BRCA1)

Goal

▶ Compile data on BRCA1 interactors & interactions, identify binding-site residues

▶ Run, adapt, and improve current framework for IDP binding prediction

▶ Develop new algorithm that characterizes and predicts IDP binding interactions

▶ Potential for disruption or blocking of protein-protein interactions

▶ Potential for synthesis of mimic regions

▶ Develop novel drugs that inhibit or replicate BRCA1 behavior

Methods: Gene Metacore and String-DB

- Databases of protein-protein interactions
- Rich ontologies for diseases and processes with hierarchical or graphic output

Results: Pathway Mapping

- BRCA1 functions independently and jointly in a complex
  - Complexes include BASC, BARD

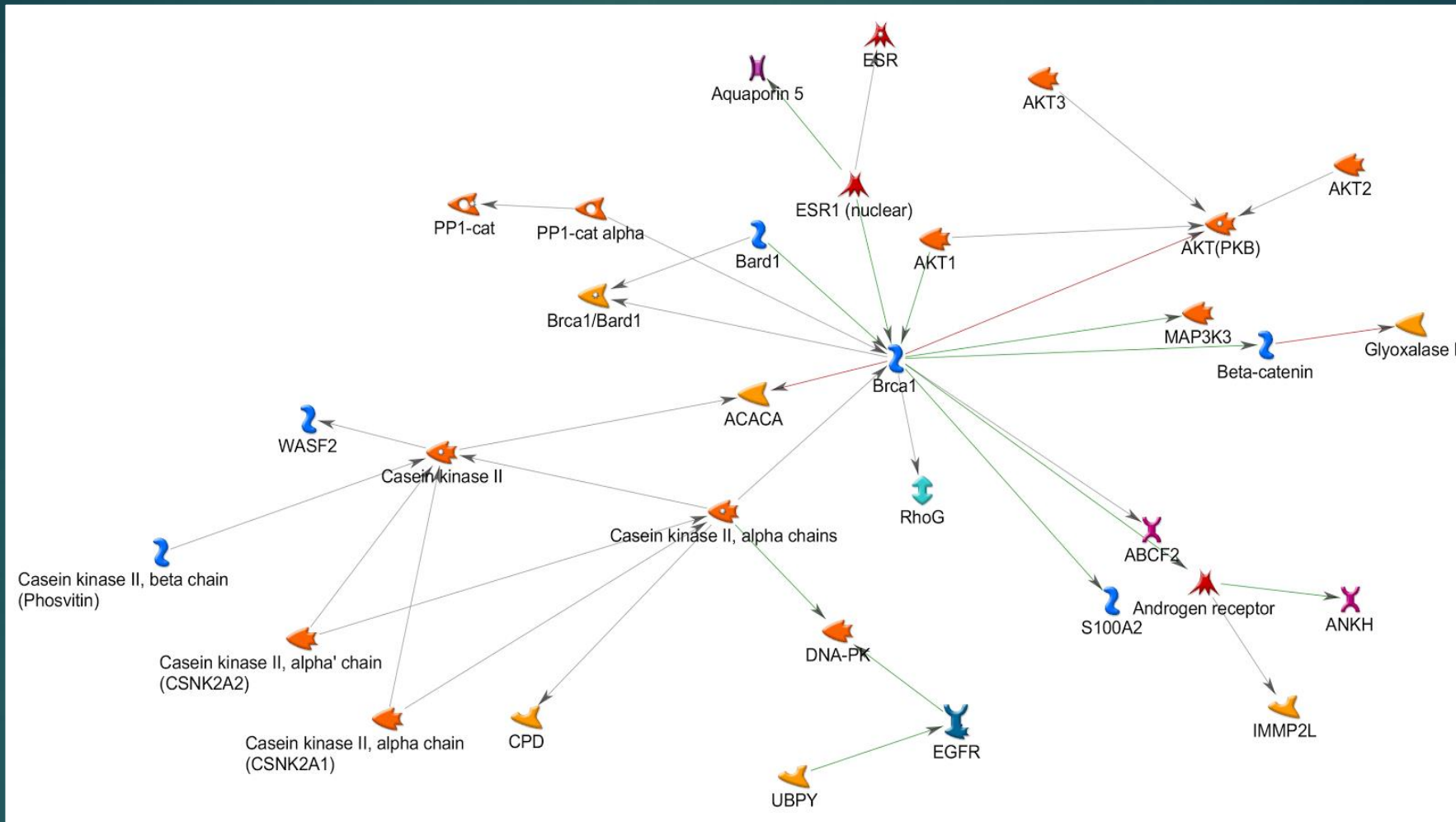- Proteins that interact with BRCA1 function independently and jointly in a complex

Figure 5: Example of BRCA1 interaction map. Image generated by Gene Metacore (shown is one map among many other pathway network maps & lists)
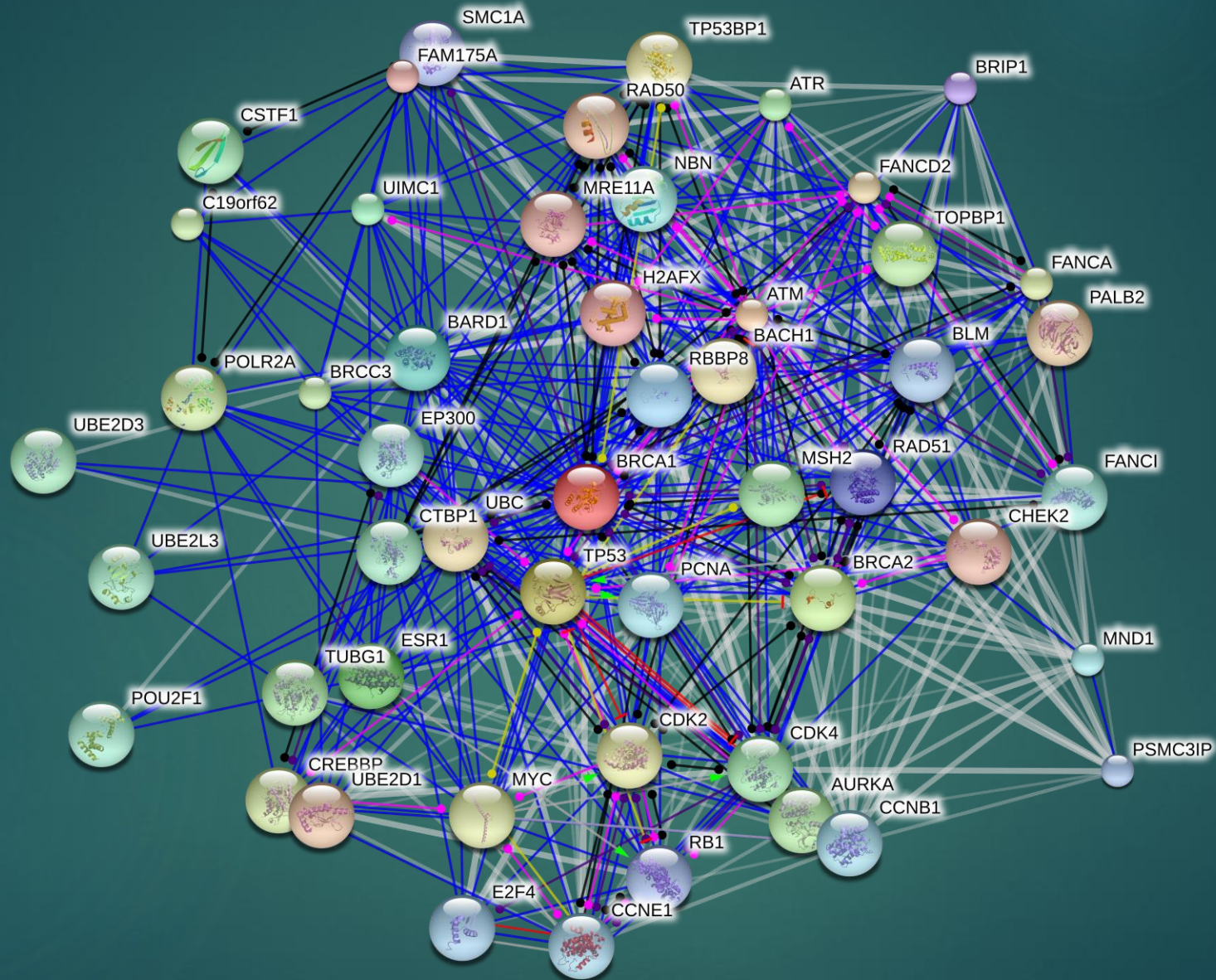
Figure 6: Example of BRCA1 interaction map. Image generated by String-DB (shown is one map among many other pathway network maps & lists)

Figure 7: Example of BRCA1 interaction table. Image generated by Gene Metacore (shown is one table among many other network data tables & lists)

Figure 8: Enrichmeny analysis on BRCA1 interactors. Data from Gene Metacore

Methods: ΔASA and DSSP

- Accessible Surface Area
  - => infer binding-site residues

- Define Secondary Structure of Protein
  - => infer secondary structure from 3D coordinates of atoms

- *Problem*: no crystallized structures of BRCA1's disordered region in Protein Data Bank

accessible surface

van der Waals surface
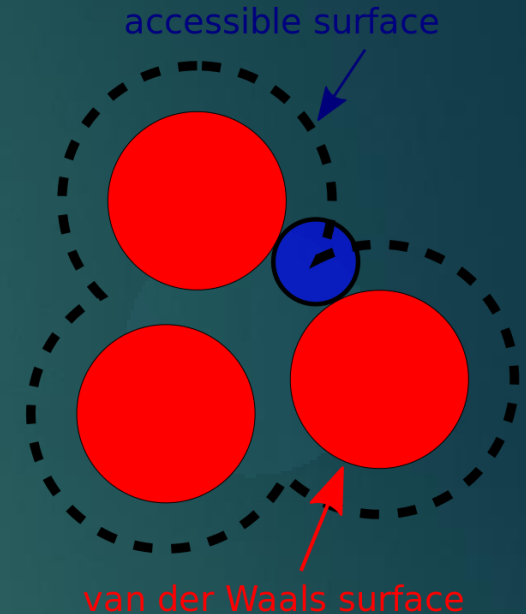
Results: ΔASA and DSSP

index	insertCode	res	buried
1649	ARG	0	
1650	MET	0	
1651	SER	0	
1652	MET	0	
1653	VAL	0	
1654	VAL	1.40794	
1655	SER	8.27484	
1656	GLY	27.5648	
1657	LEU	19.211	
1658	THR	8.9448	
1659	PRO	7.01606	
1660	GLU	0	
1661	GLU	0	
1662	PHE	0	
1663	MET	0	
1664	LEU	0	
1665	VAL	0	
1666	TYR	0	
1667	LYS	0	
1668	PHE	0	
1669	ALA	0	
1670	ARG	0	
1671	LYS	0	
1672	HIS	0	
1673	HIS	0	
1674	ILE	0	
1675	THR	0	
1676	LEU	0	
1677	THR	0	

Figure 9: ΔASA data on 1t15 model on PDB. Calculations conducted on every model of BRCA1 in PDB.

Methods: Machine Learning

- MoRF: Molecular Recognition Feature
- Given MoRF, categorize potential binding sites in terms of class, fold, family, and superfamily and provide a yes/no binding prediction
- Database of 482 MoRFs
- Bayesian network learning
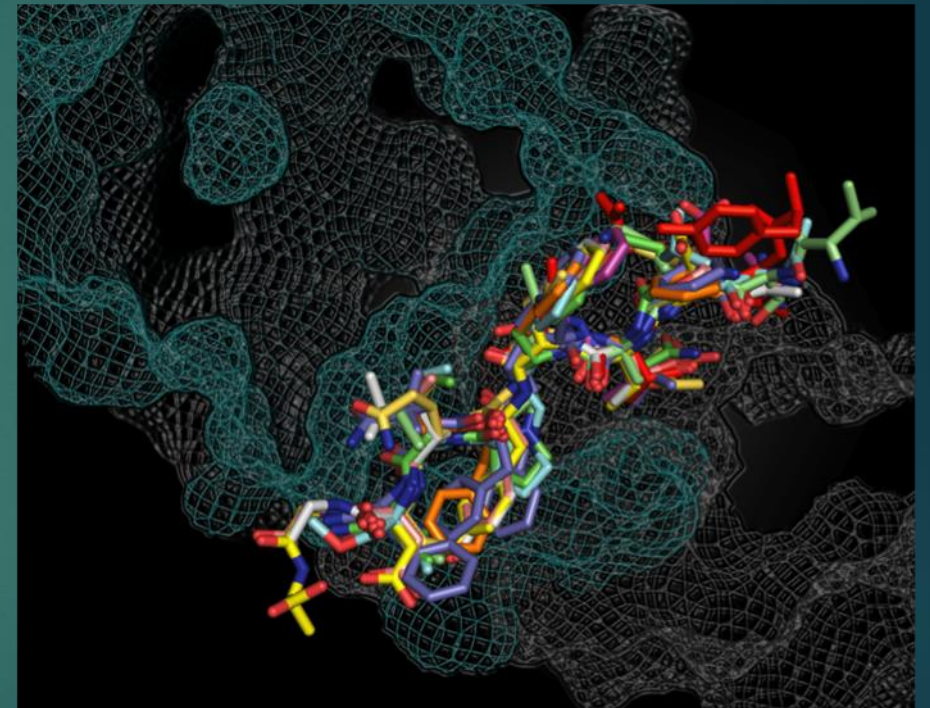
Figure 10: HIV MoRF (Jonah Kallenbach, personal communication)

Results: Machine Learning

Figure 11: KNIME Workflow adapted from Jonah Kallenbach's research at RSI

Methods: Homology

- ▶ Correlate IDPs with other proteins
- ▶ Infer function of same residues on BRCA1
- ▶ BLAST: Basic Local Alignment Search Tool
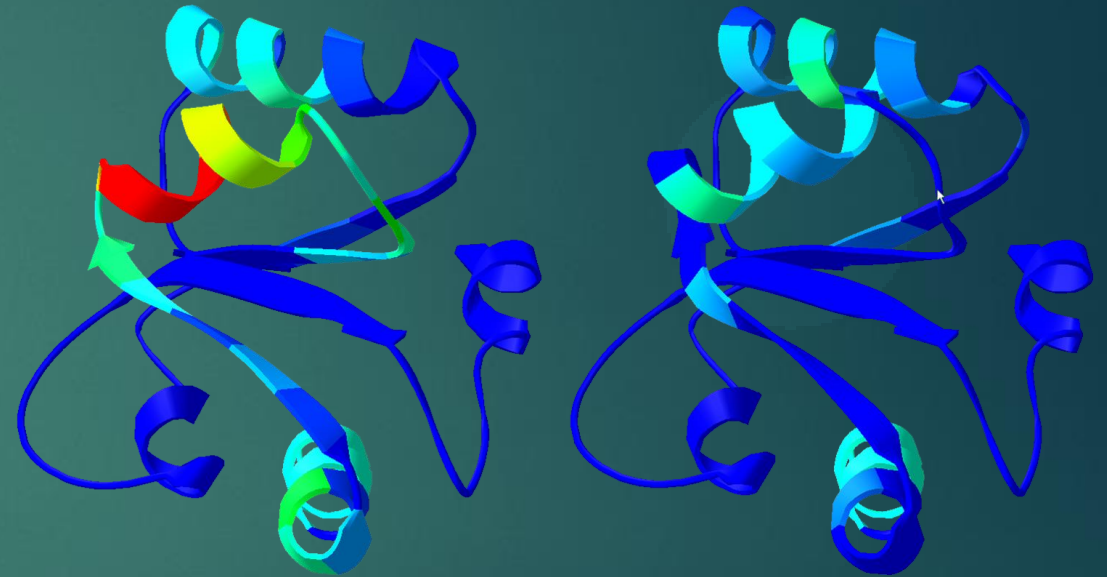
www.rscb.org
Figure 13: Example of proteins with similar sequences

Ongoing Work and Future

▶ Draw subnetworks of BRCA1 functional network based on protein function and results of enrichment analysis

▶ Developing a more accurate prediction algorithm based on

  ▶ Binding-site residue data and MoRF data

  ▶ Interaction data from databases and enrichment analyses

  ▶ Tailor to BRCA1

Future

- Validation by experts at Dana-Farber Cancer Institute

- Development of drugs that replicate BRCA1 behavior by imitating MoRFs and binding residues

Acknowledgements

- MIT PRIMES for this opportunity
- Dr. Gil Alterovitz for his guidance
- Mr. Mingjie Chen for his guidance
- Ms. Wei-lun Hsu for her guidance
- Mentor group
- Parents

Questions

Extra: Bayesian Networks

- Directed acyclic graph, probabilistic graphical model
- Nodes: random variables
- Edges: conditional probabilities

```
        Sprinkler  ←──────  Rain
            \                /
             \              /
              ↘            ↙
               Grass Wet
```