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

www.rscb.org
Figure 1: PDB rendering based on model 1jm7
Clinical Significance: BRCA Mutations and Cancer

- BRCA mutation in 90% of familial breast and ovarian cancers
- BRCA mutation $\rightarrow 5x$ breast cancer risk, $30x$ ovarian cancer risk
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

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: Enrichment analysis on BRCA1 interactors. Data from Gene Metacore
Methods: ∆ASA 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
Results: ΔASA and DSSP

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

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
Extra: Bayesian Networks

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

Diagram:
- Sprinkler
- Rain
- Grass Wet