Published in final edited form as:

Methods Mol Biol. 2009 ; 541: 449–461. doi:10.1007/978-1-59745-243-4_19.

Connecting protein interaction data, mutations and disease using bioinformatics

Jake Yue Chen, PhD,

Assistant Professor of Informatics, Informatics and Technology Complex (IT), Room #493,
Indianapolis, IN 46202, Phone:(317) 278-7604

Eunseog Youn, PhD, and

Assistant Professor, Department of Computer Science, Texas Tech University, PO Box 43104,
Lubbock, TX 79409, Phone:(806) 742-3527

Sean David Mooney, PhD

Associate Professor and Director of Bioinformatics Core, The Buck Institute for Age Research, 8001
Redwood Blvd., Novato, CA 94945; Phone 415-209-2038

Jake Yue Chen: jakechen@iupui.edu; Eunseog Youn: eun.youn@ttu.edu; Sean David Mooney:
mooney@compbio.iupui.edu

Summary

Understanding how mutations lead to changes in protein function and/or protein interaction is critical to understanding the molecular causes of clinical phenotypes. In this method, we present a path toward integration of protein interaction data and mutation data, and then demonstrate the identification of a subset of proteins and interactions that are important to a particular disease. We then build a statistical model of disease mutations in this disease-associated subset of proteins, and we visualize these results. Using Alzheimer's disease (AD) as case implementation, we find that we are able to identify a subset of proteins involved in AD and we are able to discriminate disease-associated mutations from SNPs in these proteins with 83% accuracy. As the molecular causes of disease become more understood, models such as these will be useful for identifying candidate variants most likely to be causative.

Keywords

Protein interaction; SNP; mutation; bioinformatics; data integration

1. Introduction

Systems approaches are critical to understanding clinical phenotypes. One important challenge for further research is building functional knowledge on how genetic variation affects the function and expression of proteins (or some other functional unit) in a biological network. Until recently, a systems understanding of the effects of variation has been an elusive challenge, and it remains a significant opportunity for research. There are several resources that have begun to address this issue. Databases such as dbSNP (1), KEGG (2), UniProt/Swiss-Prot (3) are beginning to include information useful for understanding the prevalence of variation in proteins, functional domains, pathways and systems. Meta-resources are also being actively developed for understanding the effects of variation on systems. SNPs3D (<http://www.snps3d.org/>), for example, displays candidate pathways to links with functional predictions of nonsynonymous SNPs (4). The Pharmacogenetics Knowledgebase (PharmGKB, <http://www.pharmgkb.org/>) has well curated pathways with links to variation submitted to the resource, and phenotype data whenever available (5).

The task of linking genetic variation to functional protein effects in a biological system is complicated by the variety of effects that genetic variation can impose on a gene or gene product. Nonsynonymous SNPs can affect protein stability and function, while synonymous SNPs can effect transcript alternative splicing; noncoding SNPs can also affect gene transcription(6). Although there are analytical tools for prediction of functional nonsynonymous variation, other types of SNPs have been difficult to quantify. Our current efforts focus on how nonsynonymous mutations affect both protein function and the interplay of these proteins in the pathway contexts. Recently, researchers have begun to perform computational studies that address these important research questions. In particular, Ye, *et al.* showed that many disease-associated mutations from Swiss-Prot are likely involved in disrupting protein interactions (7). They also found that Swiss-Prot mutations were distributed differently than SNPs using comparative protein structural models (7). PharmGKB has also been curating pathways and genotype data associated with pharmacogenetics (5).

A continuing open research question today is how to relate genetic mutation data to protein interaction networks, and therefore to the molecular causes of disease. Addressing this question raises several challenges. First, different types of global data sets must be integrated so that interaction data and mutation data can be arbitrarily queried, visualized, and analyzed together. Second, the disease context of the mutations and the proteins within the system must be revealed and understood. Finally, it would be useful to have a model for how mutations cause disease within the specific context of protein interaction sub-networks. However, there has not been immediate solutions to these challenges, although significant advancements have been made (4,7). Further complicating this is that databases of mutations tend to be highly biased towards well-characterized proteins of interest, and therefore skewing the distributions of mutations compared to natural occurrence. Similarly, nonsynonymous SNPs may not be neutral and may alter the function of proteins or disease-associated mutations may only be in linkage or linkage disequilibrium with the causative allele (depending on the the genetic approach).

We present a method that explores solutions to the challenging questions of connecting genetic mutation data effects in interaction networks. Our initial case study uses a database of human protein interactions and mapped annotated SNPs among interacting proteins, which are involved in the Alzheimer's Disease (AD) process. To this end we built a database consisting of all related protein interactions, participating protein's "AD relevant" score, and all annotated mutations and SNPs. We built a model for disease-causing mutations from a specific set of interacting proteins that are associated with the AD phenotype. Using a model trained on the most highly ranked AD proteins, we are able to discriminate phenotypically associated mutations from polymorphisms with nearly 83.7% accuracy. Interestingly, the model was more accurate when trained on the subset of proteins most likely to be associated with AD, while the full set of proteins resulted in poorer performance. When we mapped the SNP data onto interacting proteins, we also observed that, while deleterious SNPs are broadly distributed among interaction hub proteins and peripheral (non-hub) proteins in the AD sub-network, most highly ranked AD proteins (subnetwork hubs) contain high ratios of deleterious SNPs, suggesting a link between the systems properties of protein functions and disease phenotypes.

We take the following steps toward integration and analysis of mutations in protein interaction data:

1. Collect a dataset of experimentally determined protein interactions, and import the interactions into a relational database.
2. Using phenotypically annotated mutation data and SNPs derived from several sources, identify nonsynonymous genetic variation for the proteins identified in step 1.

3. Identify proteins relevant to the specific phenotype of interest by identifying proteins highly connected with known disease-associated proteins.
4. Build a matrix of sequence and structure-based features for each mutation in the subset of proteins identified in step 3.
5. Train a supervised learning algorithm to discriminate disease-associated mutations from polymorphisms using the matrix from step 4.
6. Visualize the model on a protein interaction network

2. Materials

2.1 Protein Interaction Data

To integrate the protein interaction data we used the following steps:

1. We obtain the human protein interaction data from the Online Predicted Human Interaction Database (OPHID) (8). OPHID is a comprehensive and integrated repository of publicly known or predicted human protein interactions, which are derived from curated literature, high-throughput experiments, or computational inference from homologous protein interactions in model organisms. Predicted human interactions are usually confirmed with additional evidence such as domain co-occurrence, co-expression, and GO semantic distances (8).
2. We used the protein identifier mapping table provided at the ftp site of the UniProt Knowledgebase (3) to perform mapping between human protein's Swiss-Prot IDs and their UniProt IDs for all OPHID proteins.
3. The entire collection of processed OPHID data were subsequently loaded into an Oracle 10g relational database system using PERL program parsers and Oracle DBMS sqldr data loading software facility.
4. For each interaction pair, we further assigned a heuristic interaction confidence score, based on a simple scoring method described in (1). The approach can be summarized as the following. First, we manually determine a confidence score between 0 and 1 (the higher, the better) that provides an estimate of data quality reliability. We assign heuristic scores of 0.9, 0.5, and 0.3 to protein interactions obtained from human, medium-quality non-human mammalian systems, and high-quality non-mamalian systems respectively. Then, we perform a score combination using the following formula, where $p_{INT}(a, b)$ is the overall confidence score for interaction between proteins a and b , and $p_i(a, b)$ is the confidence score for a specific interaction derived

$$\text{from resource } i \text{ between proteins } a \text{ and } b: p_{INT}(a, b) = 1 - \prod_i (1 - p_i(a, b))$$

For the case study described here, OPHID was downloaded in February 2006. This release of OPHID contains more than 47,213 human protein interactions among 10,577 proteins identified by Swiss-Prot IDs. After the mapping to UniProt, the database contains 46,556 unique protein interactions among 9959 proteins identified by UniProt IDs.

2.2 Data Integration

We used Oracle 10g as the data integration platform and developed a coherent relational database to integrate data from protein interactions, SNPs, mutations and functional predictions, and the AD-specific disease protein list into relational database tables. The data integration process requires the use of PERL scripts to parse data generated from different sources, and SQL scripts to join different tables together to format the data into final report views.

We obtained mutation and SNP data from several sources. For bulk SNP data, dbSNP is the central repository (<http://www.ncbi.nlm.nih.gov/projects/SNP/>), and is distributed in parseable XML format. Nonsynonymous SNPs are annotated with protein accessions and positions. For disease-associated mutation data, several options are available. First, Swiss-Prot (<http://www.expasy.org/>) entries contain mutations in the VARIANT records. Second, OMIM (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>) contains mutations associated with disease with natural language phenotype annotations. Other options include the semi-private HGMD (<http://www.hgmd.org/>) and locus specific databases.

Here we describe our process to extract mutation and SNP data from Swiss-Prot files:

1. Mutation and SNP lists were determined by downloading each ID in UniProtKB/Swiss-Prot format using a BioPerl (<http://www.bioperl.org/>) script that utilizes the Bio::DB::SwissProt module.
2. The resulting files were then parsed using a home-grown Python script, by extracting the feature tags from each protein. Note that although we used an informal script to parse the Swiss-Prot entries, many users will likely use Swissknife (<http://swissknife.sourceforge.net/>), an object oriented Perl library to handle these files.
3. When parsing the feature tags, we only considered VARIANT features and did not consider mutagens and VAR_SEQ features. dbSNP entries are annotated with a dbSNP rsid and mutations are annotated with a phenotype or left unannotated. Some SNPs contain both a phenotype annotation and an rsid. Swiss-Prot was chosen because it is generally well annotated and contains both mutation and SNPs within the same context.

3. Methods

3.1 Identification of proteins associated with a specific disease or phenotype

The identification of disease-specific proteins was done in the following steps:

1. To obtain a list of disease specific proteins, we performed a search of the OMIM database, retrieving each OMIM gene record in which the “description” field contains the term “Alzheimer”.
2. These initial proteins were used as “seeds” to expand into a interaction data set to proteins with high confidence interactions, described in detail in (1).
3. Using the database built in **Materials 2.2**, all Swiss-Prot SNPs and mutations were identified in the expanded set of proteins.

For the AD related case study, 65 OMIM genes or 70 proteins (identified by Swiss-Prot IDs or UniProt IDs) were retrieved from the initial search on OMIM. These 70 proteins were used as “seeds” to expand into a interaction data set (AD interaction subnetwork) consisting of 655 expanded proteins as described in detail in (1). It should be noted that the full set of interacting proteins contains all possible interactions, not necessarily an interaction important, or even observed, in AD tissues. For each of 70 AD proteins, mutations from Swiss-Prot were extracted as described in Materials. Among the 655 total AD expanded interacting proteins identified, 240 proteins have been found to contain 2,941 mutations, which were annotated by the Swiss-Prot database with information from the dbSNP with various annotated phenotypes. In Figure 1, we show a histogram including all of the number of dbSNP and phenotype mutations by protein. Annotated mutations ranged from 0 to a maximum of 210 in the androgen receptor (ANDR_HUMAN). Other proteins highly populated with mutations include TP53_HUMAN (193 mutations), FA8_HUMAN (185), CO4A5_HUMAN (141) and FA9_HUMAN (138).

3.2 SNP Function Prediction

Initially, methods for SNP function prediction were based on conservation in multiple sequence alignments derived from similarity searches against nonredundant sequence databases and the mutations could be either ranked or scored based on parameterization. Parameterization was either performed using unbiased mutagenesis experiments in LacI, T4 Lysozyme and/or HIV protease, such as SIFT (9), or using structure/function rules, such as PolyPhen (10). SIFT is used as a feature here and SIFT based predictions can be determined by running the method on their website (<http://blocks.fhcrc.org/sift/SIFT.html>). PolyPhen was not used here, but can be run from the website (<http://coot.embl.de/PolyPhen/>).

Later, more sophisticated approaches were utilized using decision trees (11) or support vector machines (12,13). Generally, the primary predictive features for such methods are derived from protein sequence, protein sequence conservation, structure, structural rules; specifically, the most important features are generally the identities of the wild type and mutant amino acids, sequence conservation and the degree of residue burial in a protein structure (11). These prediction methods generally require the use of a statistical modeling package such as R or MATLAB.

3.3 Model to predict mutations within the context of disease

To describe how we build a support vector machine (SVM) model for the mutations in a set of proteins, we will walk through the previously described AD case study. SVMs were chosen because they have been applied to mutation data previously with good performance (4,14–16). The goal of such a model is to discriminate disease-causing mutations from polymorphisms. It is important to understand that the set of proteins in question can be derived from several sources. For example, they can be a database of proteins with annotated SNPs or mutations, a family of related proteins, or functionally connected proteins in a protein interaction network.

To build a model we employ the following steps:

1. Using the previously determined set of proteins associated with a specific phenotype, the SNPs and mutations associated with each protein are determined (see **2.2 Data Integration**).
2. Each of the mutations and SNPs in the set based on protein interactions is annotated with features based on four different attributes. They include the p-value score from SIFT (9), sequence conservation score (based on information theory), and a window of sequence neighbors and blosum62 substitution score (17). Sequence conservation was evaluated using a position specific scoring matrix (PSSM) from the output of PSI-BLAST (18). PSI-BLAST is performed using the BLASTPGP (<http://www.ncbi.nlm.nih.gov/BLAST/download.shtml>) package from NCBI and is queried against the nonredundant protein database (NR) from GenBank (<ftp://ftp.ncbi.nih.gov/blast/db>). Three scores from the output were extracted; a PSSM, a weighted observed percentage, and information per position score for each residue in the sequence. For the sequence conservation of adjacent residues, a window size of 20 using the information per position score was considered. For N-terminal and C-terminal residues, we used an average conservation score of the sequence to fill in the N-terminal (C-terminal) residue scores. For the window of residue type, we considered window of size 21; 10 residues on the N-terminal side (left), 10 residues on the C-terminal side (right) and itself. Our description of the window of residue type has a similar protocol to those published for classifying protein phosphorylation sites (19). Other features are possible, including those based on sequence-based data

mining tools and protein-structure based approaches; see **Notes** (Section 4) for more information.

3. Each mutation is labeled (−1: dbSNP, 1: Swiss-Prot) and positions annotated as being in both are removed, because their ambiguous annotation prevents them from being accurately classified as neutral or damaging. There are 45 features, which are (listed by feature index id followed by feature type): 1 SIFT, 2:4 PSSM, weighted, information per position, 5:24 information per position window, 25:44 sequence window, 45 BLOSUM62.
4. For SVM classification (20), we use a linear kernel and default regularization parameter (C). We employed an SVM for classification using the SVM^{light} (21) and its Matlab interface. Since each feature has a different scale, all examples were normalized to the [0, 1] interval. The ratio between deleterious and neutral mutations is 13 : 1. To overcome this imbalanced data training, we gave more weight to negative (neutral) samples. All evaluations were performed using leave one-out cross validation.

To continue the example of AD related proteins, the 655 AD related protein in the protein interaction subnetwork were determined by looking at all interactions with the AD proteins which were extracted from OMIM. Second, we have the subset of the 70 most highly ranked proteins, based on the previously described analysis. SIFT was run on all 655 proteins giving a total 2,893 annotations with an average SIFT p-value score of 0.118. When separated by dbSNP, the phenotypically annotated mutations had an average p-value of 0.095, while the SNPs in this set had a score of 0.290. Similarly the subset of 70 important proteins (35 of which actually contained mutations) had both disease annotated mutations and SNPs. The disease annotated mutations had a SIFT p-value average of 0.110 and the SNPs had an average p-value of 0.327. Similar trends have been observed before, because SNPs are less likely to affect protein function than disease-associated mutations. We then encoded the 367 mutations annotated in the subset of proteins using the features described above.

As an example, we compare the performance differences between protein SVM predictions and protein SIFT predictions. Not surprisingly, the SVM method gives improved performance over the SIFT method alone, although SIFT remains a highly valuable feature (Table 1). Overall the model has 83.7% accuracy and is more sensitive than SIFT. Interestingly, when the model is applied to an expanded set of interacting proteins less likely to be related to AD, the model's performance declines to a level similar to SIFT (accuracy of 70.94% versus 71.75%).

3.4 Feature selection

We then used feature selection to find a subset of features for optimal classification. An important part of feature selection is to rank features according to their importance for class discrimination. Feature ranking is based on the weight associated with each feature in a classifier and can be determined by recursive feature elimination (RFE). RFE method computes the feature ranking as follows:

1. The SVM is trained and a weight associated with each feature is computed using all of the features initially.
2. The feature with the smallest magnitude of the weight is removed. This feature is the least important one. Leaving this feature out, we retrained the SVM and recomputed all the weights.
3. Steps 1–2 are iterated until all of the features are exhausted. In this way, features are recursively eliminated and ranked. The feature eliminated last is the most important

feature. We implemented the RFE feature ranking method in Matlab. RFE with an SVM was described in Guyon's paper on microarray datasets (22).

We wanted to evaluate which features were more important than others for class discrimination. The top ranked features based on the SVM-RFE method for the AD related protein set include SIFT, and information per position from PSI-BLAST for the central and adjacent residues. Based on this feature ranking, we computed performance using top ranked features (Table 2). The highest performance is achieved using the top 25 or 30 features.

3.5 Data Visualization

Due to the large scale data integration nature of this study, we chose network data visualization software with the following two features, 1) ability to directly support relational database queries as input to the visualization environment; 2) ability to combine network features, scores, and data variables created in real time with maximal flexibility for the final visualization output. Therefore, we chose ProteoLens (<http://bio.informatics.iupui.edu/teolens>, manuscript in preparation) over the popular software tools such as Cytoscape (23). Using ProteoLens, we are writing SQL queries to generate network data and prepare annotated "data associations" either as "node associations" or "edge associations". Examples of the "node associations" are protein target score identified by protein ID and the protein's gene deleterious SNP count identified by proteins. Examples of the "edge associations" are protein interaction pairs identified by both protein A ID and protein B ID and protein interaction confidence score identified the same way. The node associations are mapped to the network node's display property such as node size, shape, and color, whereas the edge associations are mapped to the network edge's display property such as edge width, color, and type, all automatically by the ProteoLens software. The final visualization can be generated as either a PDF file or a PNG image file, specified by the user.

In Figure 2, we show a visualization of the AD protein interaction subnetwork, in which protein interaction, SNP annotation, and disease protein curation information is integrated. Note that most protein interactions are connected above statistical chance to form a large connected subnetwork, which underlies the context of the AD biological process. When we only focus our attention onto the subnetwork highly relevant proteins with matched SNP annotations (large nodes with non-default colors as shown in the figure), we do see a correlation between the protein's relevance to the AD process and the percentage of deleterious SNPs for those proteins. We can also observe other phenotypical effects among non-AD relevant proteins (seen as peripheral yet red-colored nodes in the figure), primarily because these proteins, although insignificant in the AD process, play important roles in other non-AD disease processes (e.g., p53, etc, and btk, which are all peripheral in the AD subnetwork but are implicated in other diseases).

4. Notes

4.1 Use of other features for the classification of SNPs and mutations

It is important to note that the features described here are only a subset of the possible features that can be used to classify mutations. Features based on protein structure information or comparative modeling information have been used previously with improved results (11,12). Features based on phylogenetic information have also been shown to be useful (12).

4.2 Choice of a neutral set of mutations to compare with disease

In this case we chose polymorphisms from Swiss-Prot as a model of a neutral set of mutations to compare against disease-associated mutations. A subset of nonsynonymous SNPs are known to functional, that is they likely alter the protein products function. A more rigorous approach

might be to use SNPs experimentally shown to be neutral or to use saturation mutagenesis experiments for comparison (11).

References

1. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Helmberg W, Kapustin Y, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E. *Nucleic Acids Res* 2006;34:D173–80. [PubMed: 16381840]
2. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. *Nucleic Acids Res* 2006;34:D354–7. [PubMed: 16381885]
3. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B. *Nucleic Acids Res* 2006;34:D187–91. [PubMed: 16381842]
4. Yue P, Melamud E, Moulton J. *BMC Bioinformatics* 2006;7:166. [PubMed: 16551372]
5. Klein TE, Altman RB. *Pharmacogenomics J* 2004;4:1. [PubMed: 14735107]
6. Mooney S. *Brief Bioinform* 2005;6:44–56. [PubMed: 15826356]
7. Ye Y, Li Z, Godzik A. *Pacific Symposium on Biocomputing* 2006;11:439–50. [PubMed: 17094259]
8. Brown KR, Jurisica I. *Bioinformatics* 2005;21:2076–82. [PubMed: 15657099]
9. Ng PC, Henikoff S. *Nucleic Acids Res* 2003;31:3812–4. [PubMed: 12824425]
10. Ramensky V, Bork P, Sunyaev S. *Nucleic Acids Res* 2002;30:3894–900. [PubMed: 12202775]
11. Saunders CT, Baker D. *J Mol Biol* 2002;322:891–901. [PubMed: 12270722]
12. Karchin R, Kelly L, Sali A. *Pac Symp Biocomput* 2005:397–408. [PubMed: 15759645]
13. Krishnan VG, Westhead DR. *Bioinformatics* 2003;19:2199–209. [PubMed: 14630648]
14. Capriotti E, Calabrese R, Casadio R. *Bioinformatics*. 2006
15. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. *Bioinformatics* 2005;21:2814–20. [PubMed: 15827081]
16. Karchin R, Monteiro AN, Tavtigian SV, Carvalho MA, Sali A. *PLoS Comput Biol* 2007;3:e26. [PubMed: 17305420]
17. Henikoff S, Henikoff JG. *Proc Natl Acad Sci U S A* 1992;89:10915–9. [PubMed: 1438297]
18. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. *Nucleic Acids Res* 1997;25:3389–402. [PubMed: 9254694]
19. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. *Nucleic Acids Res* 2004;32:1037–49. [PubMed: 14960716]
20. Vapnik, VN. *The Nature of Statistical Learning Theory*. Springer Verlag; New York, NY: 1995.
21. Joachims, T. *Learning to classify text using support vector machines: methods, theory, and algorithms*. Kluwer Academic Publishers; 2002.
22. Guyon I, Weston J, Barnhill S, Vapnik V. *Machine Learning* 2002;46:389–422.
23. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. *Genome Res* 2003;13:2498–504. [PubMed: 14597658]

Figure 1. Distribution of annotated mutations and SNPs by proteins shown to interact with an alzheimer's disease related protein

Each of the 655 identified proteins in Swiss-Prot had their mutations and SNPs cataloged from the VARIANT features in the Swiss-Prot flat file. Not surprisingly, few proteins have more than 10 SNPs while many proteins have large numbers of mutations..

Figure 2. Visualization of the Alzheimer's Disease protein interaction subnetwork

Proteins are represented as nodes with the default color of light blue and protein interactions are represented as edges between nodes with default color of light pink and line width of 1 (faintly visible in some parts of the network). The size of the nodes are drawn in proportion to the protein's subnetwork significance score (approximately subnetwork protein centrality measure) described in (1). The color of the nodes, when not shown as the default color (missing SNP information from data integration effort), is on a sliding red scale of which the color intensity is proportional to the ratio of deleterious mutations over all mutations recorded for the proteins annotated. The widths and color intensity of the edges is proportional to the level of confidence (ranging from 0 to 1) for the given protein interaction.

Table 1

Performance comparison between SVM and SIFT score predictions on the set of 70 AD associated proteins. The

	Sensitivity (%)	Precision (%)	Specificity (%)	Accuracy (%)
SVM prediction	85.5	95.5	69.1	83.7
SIFT prediction	70.8	94.3	66.7	70.3

Table 2Performance comparison using top k features by SVM prediction.

Top k features	Sensitivity (%)	Precision (%)	Specificity (%)	Accuracy (%)
5	73.9	98.0	88.1	75.5
10	80.3	97.8	85.7	80.9
15	85.2	97.5	83.3	85.0
20	84.6	97.5	83.3	84.5
25	85.9	97.2	81.0	85.3
30	85.9	97.2	81.0	85.3
35	85.2	96.5	76.2	84.2
40	85.2	95.9	71.4	83.7
45	85.5	95.5	69.1	83.7