Priority Based Functional Group Identification of Organic Molecules Using Machine Learning

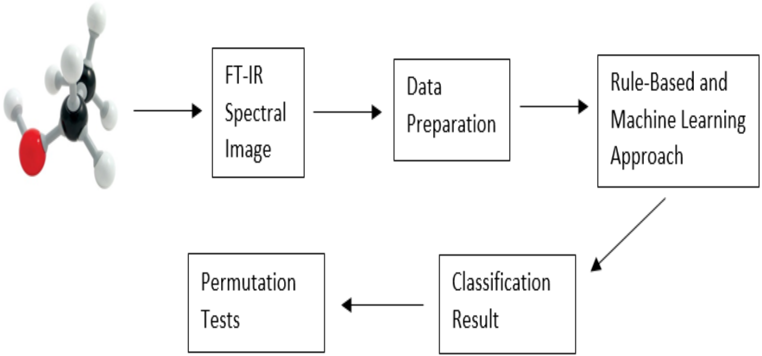
Rushikesh Nalla, Rajdeep Pinge,
Manish Narwaria, Bhaskar Chaudhury

Dhirubhai Ambani Institute of Information and Communication Technology
DA-IICT

13th January, 2018

Problem Definition

Given an organic compound, determine the dominant functional group present in the compound.



FT-IR Spectral Image

Data Preparation

Rule-Based and Machine Learning Approach

Classification Result

Permutation Tests

Motivation

- Important things to know when studying organic compounds:
  - Chemical Properties of organic compounds.
  - Types of chemical bonds present.
  - Reactivity of compounds.
- Functional groups - determine the properties of compounds
  - Eg.: Acid, Alcohol, Cyanide
- Dominant Functional Group
  - Multiple Functional Groups in compounds.
  - Chemical Properties of an organic compound mainly determined by the dominant group.
  - Less affected by the presence of other groups.
  - Dominant functional group should be the focus.
  - Consider functional group priority order.

Functional Group Priority order

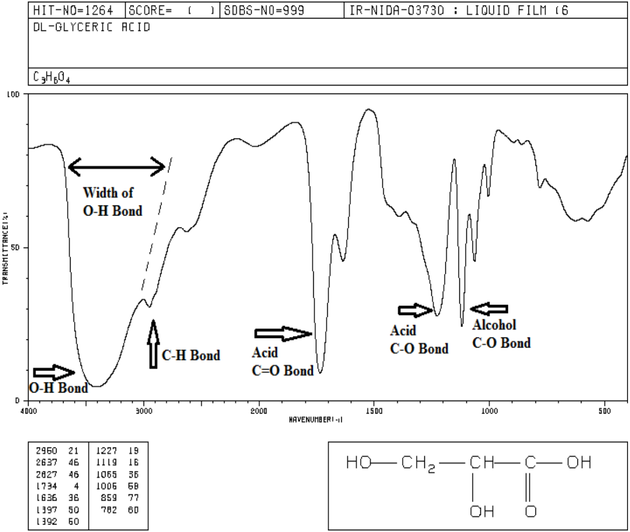
Priority Order of Functional Groups

	Functional Group	Prefix	Suffix
1	Carboxylic Acid	carboxy-	-carboxylic acid -oic acid
2	Ester	(R)-oxycarbonyl	-oate
3	Acid Halide	halocarbonyl-	-oyl halide
4	Amide	carbonyl-	-carboxamide -amide
5	Nitrile	cyano-	-nitrile
6	Aldehyde	formyl-	-al -carbaldehde
7	Ketone	oxo-	-one
8	Alcohol	hydroxy-	-ol
9	Thiol	mercapto-	-thiol
10	Amine	amino-	-amine
11	Arene (cyclic arrays of C=C)	-	benzene
12	Alkene	alkenyl	-ene
13	Alkyne	alkynyl	-yne
14	Alkane	alkyl	-ane
15	Ether	alkoxy	-ane
16	Alkyl Halide	halo-	-ane
17	Nitro	nitro-	-ane

Chemistry Approaches

1. Conduct specific chemical tests
2. Observe Fourier Transform Infrared (FT-IR) Spectrum of molecule
   - Existing approach and most common method
   - IR light is incident on the organic molecule which results in molecular vibrations
   - IR light of similar frequency is absorbed.
   - Detection process is in time domain. Converted to Frequency Domain using Fourier Transform. Hence FT-IR spectroscopy.

Visual Identification Process



HIT-NO=1264 | SCORE= ( ) | SDBS-NO=999 | IR-NIDA-03730 : LIQUID FILM (6

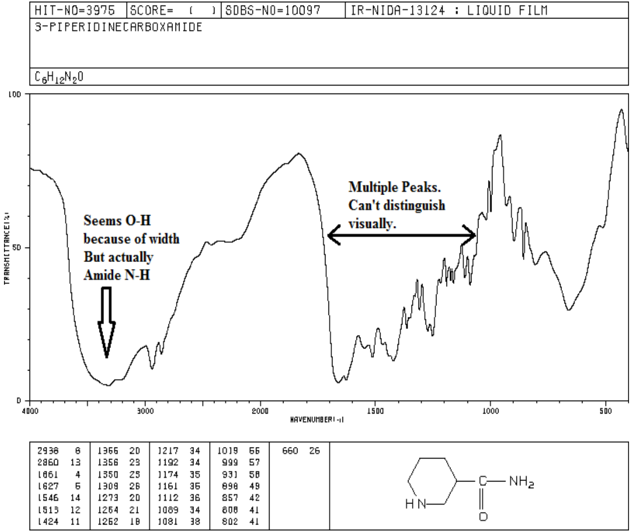
DL-GLYCERIC ACID

$C_3H_6O_4$

2960	21	1227	19
2637	46	1119	16
2627	46	1065	36
1734	4	1006	68
1636	36	859	77
1397	50	782	60
1392	60		

Figure: The visual identification of functional groups is possible due to simplicity of spectrum.

Visual Identification Process



HIT-NO=3975 | SCORE= ( ) | SDBS-NO=10097 | IR-NIDA-13124 : LIQUID FILM

3-PIPERIDINECARBOXAMIDE

$C_6H_{12}N_2O$

Seems O-H because of width But actually Amide N-H

Multiple Peaks. Can't distinguish visually.

2938	8	1366	20	1217	34	1019	56	660	26
2860	13	1356	23	1192	34	999	57		
1661	4	1350	25	1174	35	931	58		
1627	6	1309	26	1161	36	898	49		
1546	14	1273	20	1112	36	857	42		
1513	12	1264	21	1089	34	808	41		
1424	11	1262	18	1081	38	802	41		

Figure: This spectrum shows that there are problems in identification due to its complex nature.

Drawbacks of visual identification approach

- Visual Identification - which functional groups are *likely* to be present
- Depends on *experience* of chemists
- Error prone
- Time Consuming

Another Approach

- Automated, Data driven approach to identify patterns.
- Chemoinformatics $\begin{cases} Machine\ Learning \\ Spectroscopy \\ Chemistry\ (domain\ knowledge) \end{cases}$
- Challenges:
  1. Which features to use?
  2. Which classification technique to use?
  3. Can the results be generalized? - many functional groups, numerous organic compounds.
- Our Work shows that,
  - This approach is less error-prone and gives much better identification accuracy than the existing approach.
  - This approach is less time consuming than the previous two approaches.

Data Collection

- 1341 Spectrum images corresponding to 14 functional groups collected from SDBS database.
- Challenges:
  - Data in image form - Process image pixel by pixel to quantify data
  - Two different scales used on x-axis - Convert the extracted data to scale uniformly over entire range.

An FT-IR Spectrum image

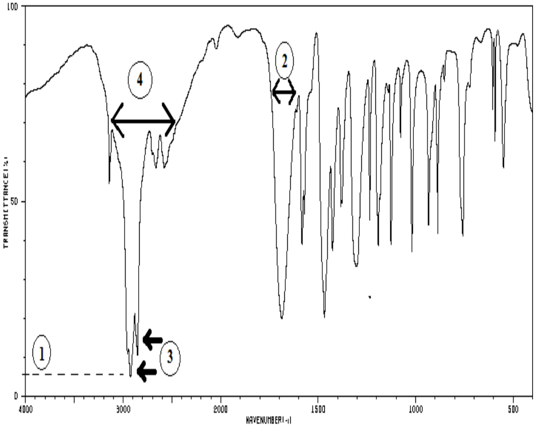


Figure: Image of FT-IR spectrum taken from SDBS database showing features used in visual identification process, (1) Transmittance Level, (2) Width of Peak, (3) Number of Peaks in the given Range and (4) Sum of Widths of All Peaks in the given Range.

Extracted Image

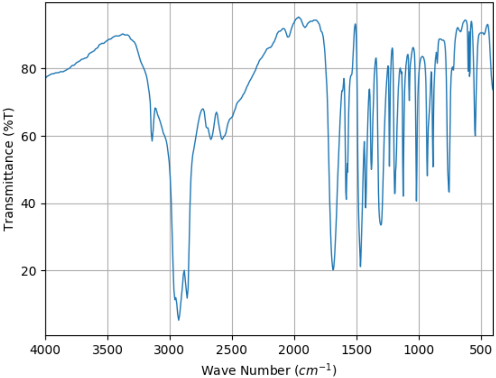


Figure: This spectrum is regenerated after extracting the data in quantified form from the original image.

Rule Based Approach

- Closer to visual identification process.
- To establish benchmark for comparison.
- Reasonably effective when the patterns in FT-IR data are distinctive and functional groups don't have overlapping features.
- Error-prone if spectra are complicated with no distinguishing features.

Rules

Standard Rules used in the visual identification process

1. Transmittance Level
2. Width of Peak
3. Number of Peaks in the given Range
4. Sum of Widths of All Peaks in the given Range

Functional Group Priority: Check for the presence of functional groups in the order of priority and terminate if such a group is found (remaining lower priority groups are not searched for)

Bond Ranges

Used to determine rules corresponding to each functional group.

Wave-number Ranges of Chemical Bonds

Bond	Wave-number
C - H	2850-2950 (Alkane) 3000-3100 (Alkene) 3290-3310 (Alkyne) 3000-3040 (Aromatic)
C = C	1620-1680 (Alkene) 1400-1620 (Aromatic)
C triple bond C	2100-2260
C = O	1690-1740 (Aldehyde) 1710-1780 (Acid) 1630-1690 (Amide) 1680-1750 (Ketone) 1735-1750 (Ester)
N - H	3300-3500 (Amine) 3100-3500 (Amide)
O - H	2500-3200 (Acid) 3200-3700 (Alcohol/Phenol)
C triple bond N	2220-2260 (Nitrile)
C - N	1025-1220 (alkyl, amine, amide) 1250-1360 (aryl)
C - O	1040-1210 (alcohols/phenols) 1210-1320 (acid)
Nitro N - O bonds	1515-1560 and 1345-1385

Rule Based Method - Results

Results of Rule Based Method

Functional Group	Total Samples	Functional Group Priority Not Considered		Functional Group Priority Considered	
		Predictions Matched	Accuracy	Predictions Matched	Accuracy
Carboxylic Acid	100	28	0.28	28	0.28
Ester	100	0	0	0	0
Amide	100	6	0.06	6	0.06
Cyanide/Nitrile	100	85	**0.85**	72	**0.72**
Aldehyde	100	50	0.5	16	0.16
Ketone	100	68	**0.68**	9	0.09
Alcohol	100	96	**0.96**	84	**0.84**
Amine	100	72	**0.72**	0	0
Arene/Aromatic	100	76	**0.76**	11	0.11
Alkene	100	73	**0.73**	14	0.14
Alkyne	49	7	0.14	0	0
Alkane	100	98	**0.98**	72	**0.72**
Ether	100	99	**0.99**	0	0
Nitro	92	32	0.35	0	0
Overall Average	**1341**	**790**	**0.5891**	**312**	**0.2327**

Data Driven Approach

- Instead of fixed rules, allow machine learning algorithm to extract the distinct patterns based on data.
- Related work done in the past: Artificial Neural Networks were explored to detect the presence or absence of individual functional groups.

Machine Learning Based Approaches

Multi-Label Approach

- Identify all the functional groups individually using some machine learning algorithm.
- Then identify the dominant functional group using the priority table.
- This approach is termed as Multi-Label Multi-Class (MLMC) Classification Approach
- Most intuitive considering the visual identification process.
- This is modification of the approach taken in past works.

Machine Learning Based Approaches

Proposed Single-Label Approaches

- For training data, adjust class labels according to priority order and give a single label to the sample
- Each sample is belongs to only one class
- Training focuses on learning only the dominant functional group pattern in the data.
- Two proposed approaches based on features used:
  1. Single Label Intermediate Approach
  2. Single-Label Multi-Class (SLMC) Classification Approach

Single Label Intermediate Approach

- Rule based Features: Transmittance Level, Width of Peak, Number of Peaks in the given Range, Sum of Widths of All Peaks in the given Range
- The above 4 features are extracted for each bond range (there are in all 23 bond ranges)
- Each sample is represented by a 92-dimensional feature vector.
- A single class label (most dominant functional group)

Multi-Label Multi-Class (MLMC) Classification Approach

- Spectroscopic data contains 3000 features per sample.
- We extract in all 250 features from the data at regular intervals.
- This is determined experimentally such that feature structure is preserved.
- Each class label is a 14-dimensional vector of 1's and 0's.

Single-Label Multi-Class (SLMC) Classification Approach

- Feature Extraction is similar to the previous approach.
- But this time there is a single class label (most dominant functional group).

Machine Learning Methods

- Multi-Layer Perceptron (MLP)
- Support Vector Machines (SVM)
- K-Nearest Neighbours (KNN)
- Random Forest Classifier (RFC)

Why these methods?

- To test the results on methods employing different learning algorithms based on varying complexity.
- The goal is not to compare machine learning methods but to test the approaches on a variety of learning algorithms.

Test Results

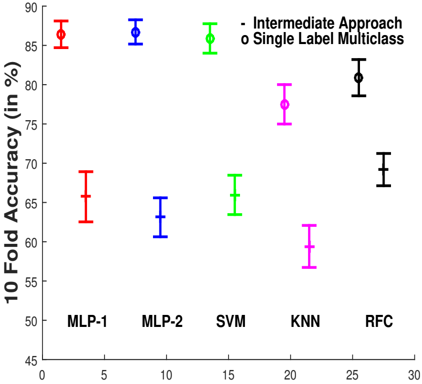


Figure: 10 Fold CV accuracies for the intermediate and SLMC methods. The error bands indicate 95% confidence intervals of the mean accuracies.

Test Results

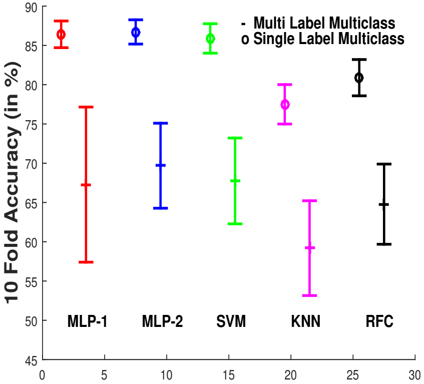


Figure: 10 Fold CV accuracies for SLMC and MLMC approaches.The error bands indicate 95% confidence intervals of the mean accuracies.

Statistical Tests

- Rule Based Approach - White Box Model (Complete transparency about rules, classification thresholds)
- Machine Learning Based Approaches - Black Box Model. (Rules, Patterns mapping samples with functional groups are unknown/hidden)
- Statistical Tests - To validate and gain insights about the ‘Black Box Model’.

Class Permutation Test

- Class labels are randomly permuted among the samples
- Thus the class structure is destroyed

Null hypothesis: Class structure is not exploited by the algorithm (no dependence between class label and features)

Results:

- p value is less than 0.01 for all algorithms.
- Rejection of null hypothesis.
- Algorithms are sensitive to class structure i.e. class label matters in classification.

Feature Permutation Test

- Features are randomly permuted within each class
- But permute features within the same feature location

Null hypothesis: Feature dependence is not exploited by the algorithm (no dependence among the features)

Feature Permutation Test

Results

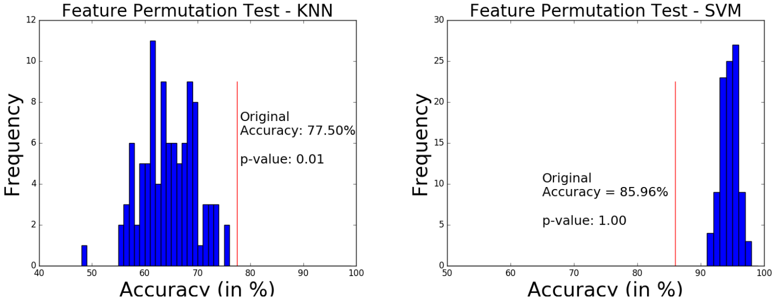


Figure: Left - the distribution of mean accuracy for KNN. Right - the distribution of mean accuracy for SVM.

Feature Permutation Test

Analysis

- Only KNN which is a simple classifier is exploiting feature dependency.
- MLP, SVM and RFC are not completely utilizing the feature dependency. This may be due to the complex nature of these algorithms.

Other Methods and Observations

- MLP with Deep Learning
- Dependence on Number of Samples
- Linear Classifier
- Principle Component Analysis (PCA)

Conclusion

- Our goal was to assign a single label which is the most dominant functional group present in the sample.
- Rule based approach gave a poor accuracy of 23.27% in priority based classification.
- The Single Label Intermediate Approach gave least accuracy (60-65%) among the three ML based approaches.
- The accuracy improved slightly (65-70%) in MLMC approach and was maximum (80-85%) for SLMC approach based on the ML method used.
- We observed that the ML approach, in general, performs better than rule based approach.
- Finally, our proposed SLMC approach works better than the traditional MLMC approach.

Future Work

- More insights into high p-values of feature permutation test.
- More experiments with white box approaches to identify groups which are performing poorly and improving their accuracy.
- Detailed analysis of current approaches like Deep Learning and PCA which have not yielded satisfactory results.
- Presenting the work for the benefit of targeted chemistry community working in this field.

References

- SDBS, National Institute of Advanced Industrial Science and Technology, Japan, `http://sdbs.db.aist.go.jp`, 2017
- Baker, Using Fourier transform IR spectroscopy to analyze biological materials, Nature Protocols, 2015
- H. Favre and W. Powell, Nomenclature of Organic Chemistry: IUPAC Recommendations and Preferred Names, 2013
- Robb, E. W. and Munk, M. E. A, Neural Network Approach to Infrared Spectrum Interpretation, 1990
- Ojala, Permutation Tests for Studying Classifier Performance, J. Mach. Learn. Res., 2010
- Duda, Pattern Classification (2nd Edition), 2000

Thank You!

Any Questions?