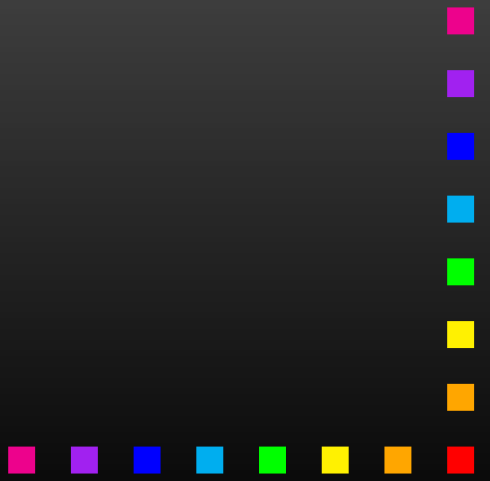
A machine learning approach for predicting the EC numbers of proteins

James Howse

Collaborators : Mike Wall, Judith Cohn, Charlie Strauss

Los Alamos National Laboratory

CCS-3 Group



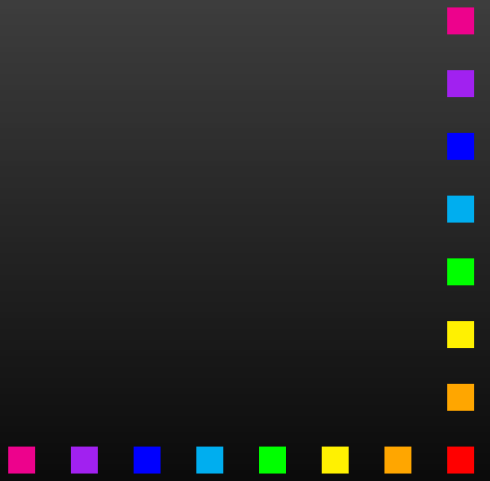
Motivation

Use sequence and/or structure similarity scores between a protein and a set of reference proteins to predict first EC numbers.

- How well does an expert predict?
- What features does the expert use?
- Can an automated system outperform the expert?
- Can an automated system approach the optimal performance?
- Does combining sequence and structure similarity produce better predictions?

Outline

- Data Sets
- Problem Description
- Reference Classifier
- Feature Space
- SVM Classifier
- Performance Comparisons
- Discussion and Conclusion



Comparison Data Sets

Protein Data ($\mathcal{D}$) : All proteins in SCOP (Version 1.65) with a single first EC number (24095 proteins).

Reference Data ($\mathcal{T}$) : All proteins in ASTRAL40 (Version 1.65) with a single first EC number (2073 proteins).

EC Labels : Supplied by EBI.

SCOP : A curated database of protein domains with known structure. Organized by structure (periodic table).

ASTRAL40 : A non-redundant subset of SCOP in which all proteins have less than 40% sequence identity

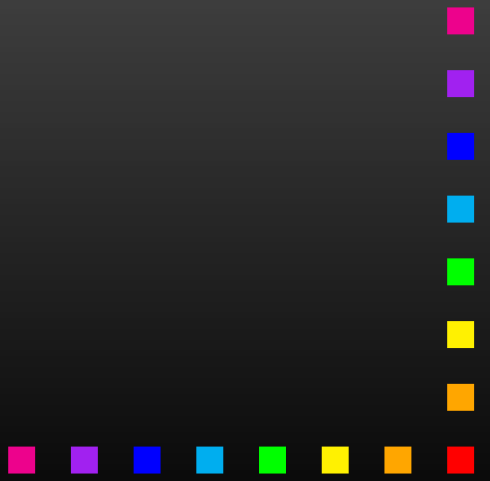
Comparison Procedure : Compare all members of $\mathcal{D}$ with all members of $\mathcal{T}$

Sequence Similarity

Tool : Psi-Blast run for 5 iterations

Transformation : Compute $e^{-E_{ij}}$ where E_{ij} is the *e-value* obtained by comparing the ith SCOP protein with the jth ASTRAL40 protein.

Range : The range of $e^{-E_{ij}}$ is $[0, 1]$ where 0 is a bad match and 1 is a good match.

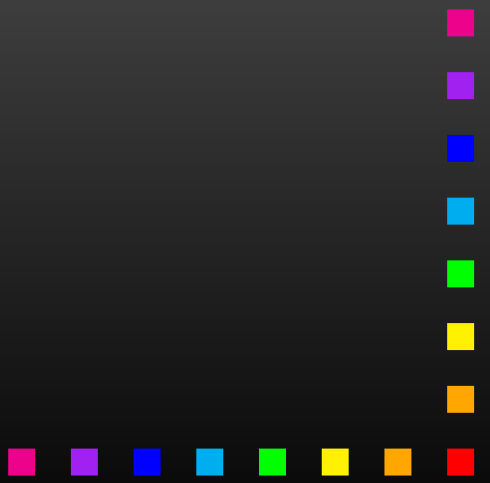


Structure Similarity

Tool : Mammoth

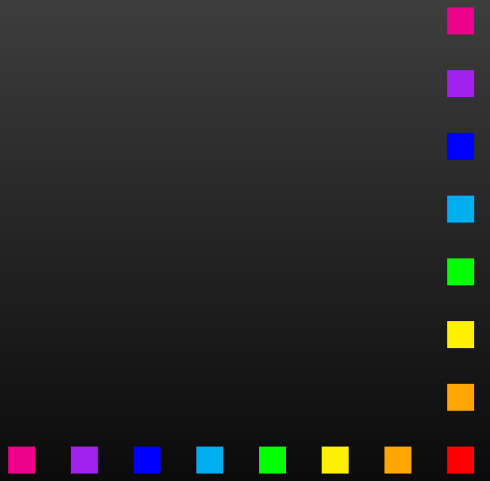
Transformation : Compute $\frac{E_{ii}}{E_{ij}}$ where E_{ii} is the *match score* obtained by comparing the ith SCOP protein to itself and E_{ij} is the *match score* obtained by comparing the ith SCOP protein with the jth ASTRAL40 protein.

Range : The range of $\frac{E_{ii}}{E_{ij}}$ is $[0, 1]$ where 0 is a bad match and 1 is a good match.



Transformation Discussion

1) Make the ranges the same
2) Make similar values represent similar match quality
3) Reduce numerical issues associated with very large or very small values



Problem Formulation

Problem : Use similarity scores to classify proteins into 1 of 6 EC categories.

Method :

- Design a 6-class classifier using similarity scores as data $\boldsymbol{x}$ and first EC numbers as labels y.
- Follow the traditional approach of selecting a feature space and designing the classifier in this space.

Performance Measure : The total error $\big(P(f(\boldsymbol{x}) \neq y)\big)$ on the 6-class problem.

Multi-class Methods

- One Versus All - Design 6 two-class classifiers where the classes are EC #k / not EC #k for $k = 1, ..., 6$.
- $+1 \rightarrow$ first EC number is k
  $-1 \rightarrow$ first EC number is *not* k
- Many simple, fast and reliable algorithms for 2-class classifier design.
- Number of required 2-class classifiers increases linearly with increasing number of classes.

Reference – Classifier

Reference is motivated by the procedure of a human expert (nearest neighbor)

1) For a protein V run a similarity comparison (sequence or structure) between V and the reference set $\mathcal{T}$.
2) Find $T \in \mathcal{T}$ such that T has the maximum similarity score.
3) Predict that the EC number of V is the EC number of T.
   a. If there are several Ts with the maximum similarity score, predict the EC number of V to be the winner of a majority vote over the EC numbers of the tied Ts.
   b. If the vote is tied, randomly choose from among the EC numbers of the tied Ts.

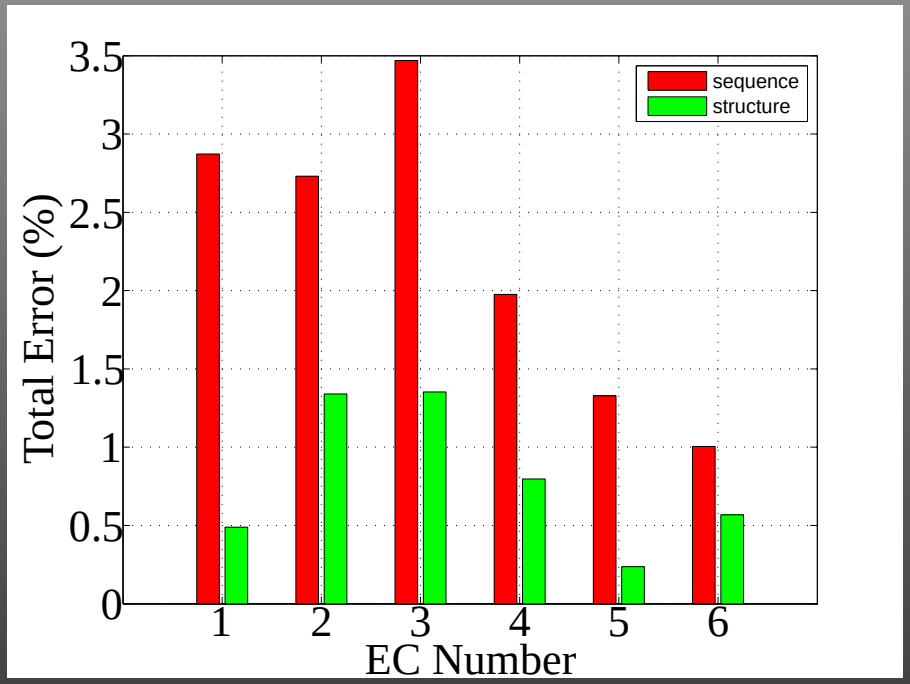
Reference – Performance

Predicting EC numbers for SCOP ($\mathcal{D}$) using ASTRAL40 ($\mathcal{T}$)

	Total Errors	Percent Error	Upper Bound (95%)	Lower Bound (95%)
Structure	567	2.353	2.545	2.162
Sequence	1647	6.835	7.154	6.517

- Computed binomial 95% confidence intervals
- With respect to 95% confidence intervals, structure has smaller error than sequence

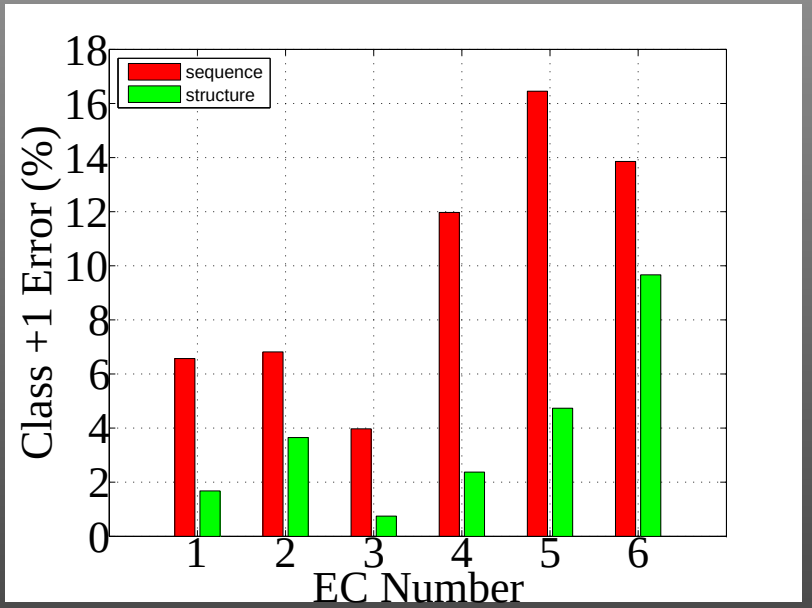
Reference – Detail

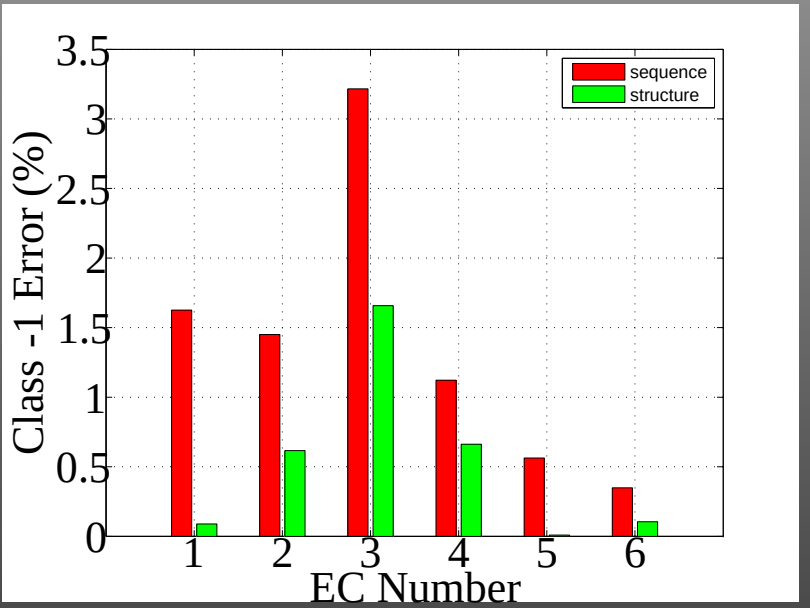


	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6
Marginal Percentage	24.36	22.56	35.01	8.56	3.68	5.81

- The reference always has smaller error than the trivial classifier.

Reference – Class Errors





- The false positive rate is much higher than the false negative rate.
- The false positive rate is generally higher for EC numbers 4,5,6 than for EC numbers 1,2,3

Feature Space – Information

Decision #1 : Choice of information

Information Expert Uses :

1) The maximum similarity score(s)
2) The EC number(s) of the protein(s) associated with the maximum similarity score(s)

Observations :

- The expert limits the similarity scores considered to those with maximum value.
- The EC numbers of the reference proteins are very important to the expert.

Feature Space – Dimension

Decision #2 : Number of dimensions

Data Dimension : Using similarity scores and EC labels for *all* proteins in the reference set $\mathcal{T}$ gives a feature space dimension d of $O(1000)$!

- Leads to poor generalization (future) performance because of the curse of dimensionality.
- Leads to large training times because computational complexity of learning increases for increasing d.

Note : Maximum number of scores p used by the expert is the maximum number of ties that occur in the maximum similarity scores, hence $p \ll O(1000)$.

Feature Space – Specification

Sequence or Structure :

1) For a protein V run a similarity comparison (sequence or structure) between V and the reference set $\mathcal{T}$.
2) Find the 25 largest similarity scores for V and sort them into descending order.
3) Multiply each score value by the label ($+1$ / -1) of the associated ASTRAL40 protein.

Sequence and Structure : Simply concatenate sequence and structure above into a $d = 50$ feature space

Note : The reference performance is unchanged.

SVM Classifiers – Method

- The primal SVM problem we solve is

$$\min_{\boldsymbol{w},b} \ \lambda \|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \max\Big[\Big(1 - y_i\big(\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_i) + b\big)\Big), 0\Big]$$

- Parametrized to normalize the problem appropriately
- Solution method obtains an ϵ-optimal solution to this *primal* problem in $O\big(n^2\big(d + \log\frac{1}{\epsilon}\big)\big)$ time
- If a property of the distribution is known, there are expressions for the relationship between λ and n
- Solution method computes appropriate values for λ and kernel parameters using a validation set

SVM Classifiers – Properties

- Error converges asymptotically ($n \to \infty$) to Bayes error e^* for *any* joint distribution P.
  - With mild assumptions on P, IID sampling is *not* necessary.
- Good finite sample rates of convergence $\left(e(f) - e^* \leq \frac{c}{n^a}, a \in (0, 1]\right)$ are obtained with mild assumptions on P.
- Convergence rates hold when classifier parameters are selected using a validation set.

SVM Classifiers – Data Sets

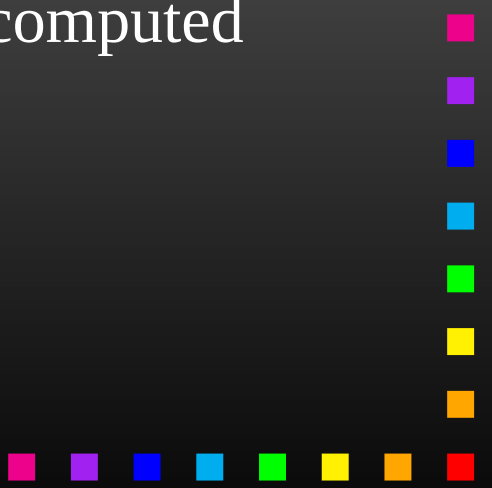
Training Set : Select 5000 SCOP proteins randomly *without* replacement.

Testing Set : Select 15000 of the remaining SCOP proteins randomly without replacement.

Validation Set : Use the remaining 4095 SCOP proteins.

Kernel : $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = e^{-\sigma \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2}$ (*not* Gaussian)

Parameters : Values for the parameters λ and σ are computed using the validation set.



SVM Classifiers – Performance

	Total Errors	Percent Error	Upper Bound (95%)	Lower Bound (95%)	Percent Change
Combined	191	1.273	1.453	1.094	—
Structure	261	1.740	1.949	1.531	-36.68
Sequence	402	2.680	2.938	2.422	-110.5

- Multi-class errors computed by using the label assigned by the 2-class classifier with the smallest discriminant value
- Computed binomial 95% confidence intervals
- With respect to 95% confidence intervals, combining decreases error

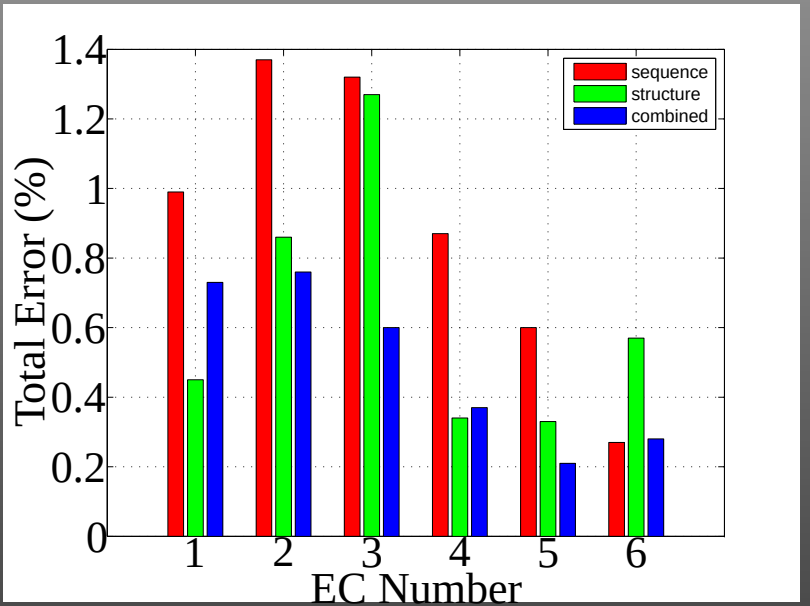
SVM Classifiers – Hypothesis Tests

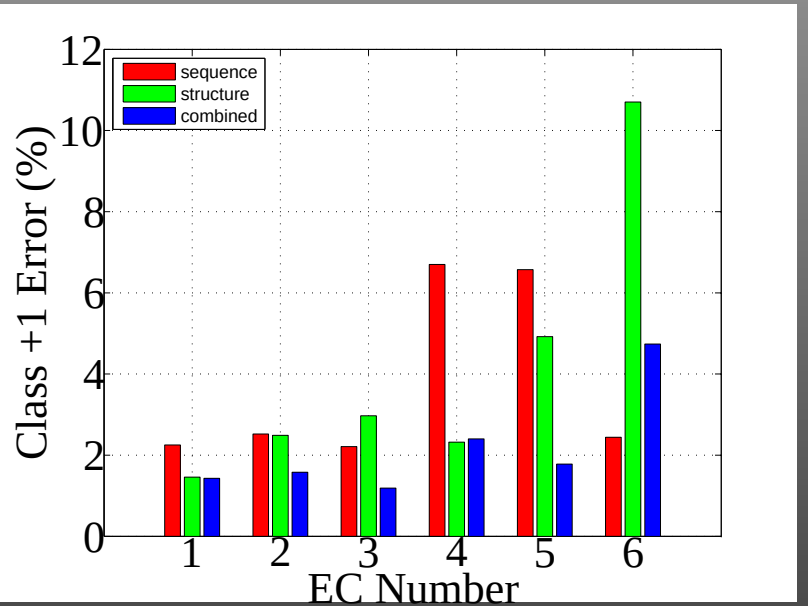
Null Hypothesis : The average error rate of the SVM classifier designed using sequence and structure is equal to the average error rate of the SVM classifier designed using sequence (structure) only.

SVM Classifier	Combined vs. Sequence	Combined vs. Structure
McNemar Statistic	141.34	20.59
Chi-Square Statistic	76.59	11.01
Confidence Threshold	$99.0\% \rightarrow 6.635$	$99.9\% \rightarrow 10.83$

Conclusion : With high confidence, the average error rate of the SVM classifier designed using sequence and structure is *not* equal to the average error rate of the SVM classifier designed using sequence (structure) only. The classifier designed using combined data is superior.

SVM Classifiers – Detail





- Even with binomial error bars, it is not clear if combining increases performance.
- Usually combining does not decrease performance.

SVM Versus Reference

	Percent Error	Upper Bound (95%)	Lower Bound (95%)	Percent Change
SVM - Combined	1.273	1.453	1.094	—
SVM - Structure	1.740	1.949	1.531	-36.68
Refer - Structure	2.353	2.545	2.162	-84.84
Refer - Sequence	6.835	7.154	6.517	-436.92

- With respect to 95% confidence intervals, SVM using sequence and structure has smaller error than either reference classifier
- SVM using structure only has smaller error than either reference classifier

Discussion – Issues

- Choosing what information to use is very important.
  - Huge dimensionality reduction in this problem
  - Importance of EC labels from the reference set
- Solving multi-class problem by combining 2-class classifiers is problematic.
- If the marginal probabilities change between test sample and future data, then future error may be *much* worse than test error.
- If error rates are small, then large data sets are needed to accurately estimate generalization error.
- A problem where a single object (protein) has multiple correct labels (EC numbers) is formally *not* a classification problem.

Conclusions

- How well does an expert predict?

  $\Rightarrow$ At best about 2.3%
- What features does the expert use?

  $\Rightarrow$ maximum similarity and associated EC number
- Can an automated system outperform the expert?

  $\Rightarrow$ With high confidence SVM has smaller error than reference
- Can an automated system approach the optimal performance?

  $\Rightarrow$ The SVM used has proven rates of convergence to Bayes error
- Does combining sequence and structure similarity produce better predictions?

  $\Rightarrow$ With high confidence combining produces smaller errors