AWS Case Study: San Francisco State University

About San Francisco State University

The Computer Science department at San Francisco State University has around 400 undergraduate students and 100 graduate students, and is engaged in both education and research. The department is currently at work on a machine learning project, called FEATURE, in collaboration with the Stanford Helix Group and supported by the National Institute of Health, (NIH Grant LM05652).

FEATURE uses machine learning to predict functional sites in proteins and other three-dimensional (3D) molecular structures. Professor Dragutin Petkovic explains: “Massively parallel optimization of machine learning involves the application of support vector machine (SVM) algorithms to thousands of training sets that are composed of hundreds of thousands of vectors. Optimal SVM parameters are found through brute-force parallelized grid searches with k-fold cross-validation. This optimization involves repeating similar operations many times independently.” Figure 1 below illustrates the FEATURE project.

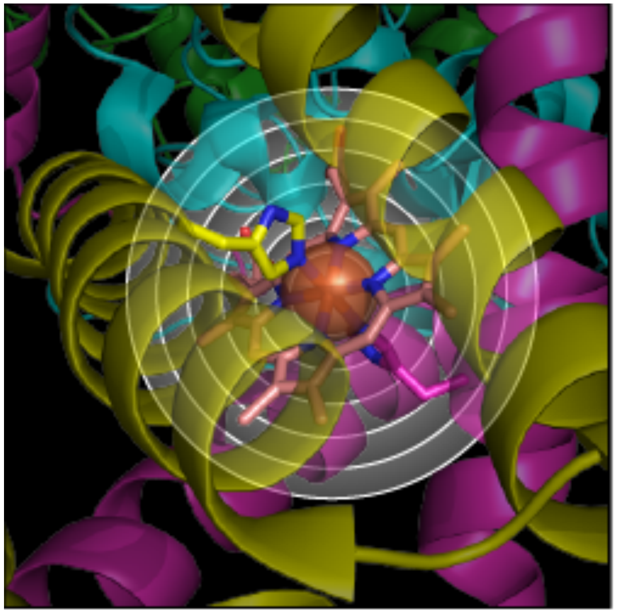
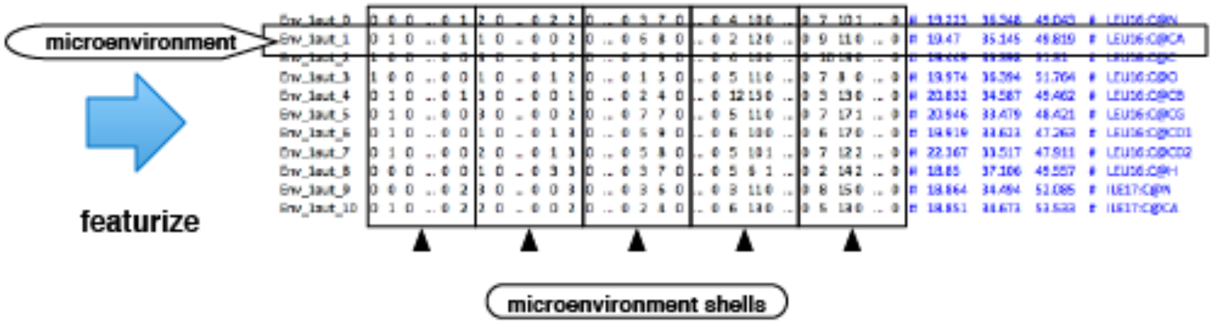


Figure A.
Conceptual diagram of a microenvironment, showing the concentric shells around a metal ion ligand.



```
Env_1aut_0  0 0 0 .. 0 1 2 0 .. 0 2 2 0 .. 0 3 7 0 .. 0 4 10 0 .. 0 7 10 1 .. 0 # 19.223 35.348 45.043 # LEU16:C@N
Env_1aut_1  0 1 0 .. 0 1 1 0 .. 0 0 2 0 .. 0 6 8 0 .. 0 2 12 0 .. 0 9 11 0 .. 0 # 19.47 35.145 45.819 # LEU16:C@CA
Env_1aut_2  1 0 0 .. 0 0 [illegible] 0 .. 0 1 [illegible] 0 .. 0 [illegible] [illegible] 0 .. 0 6 10 0 .. 0 10 14 0 .. 0 # [illegible] [illegible] [illegible] # LEU16:C@C
Env_1aut_3  1 0 0 .. 0 0 1 0 .. 0 1 2 0 .. 0 1 5 0 .. 0 5 11 0 .. 0 7 8 0 .. 0 # 19.974 36.394 51.764 # LEU16:C@O
Env_1aut_4  0 1 0 .. 0 1 3 0 .. 0 0 1 0 .. 0 2 4 0 .. 0 12 15 0 .. 0 3 13 0 .. 0 # 20.832 34.587 45.462 # LEU16:C@CB
Env_1aut_5  0 1 0 .. 0 0 3 0 .. 0 0 2 0 .. 0 7 7 0 .. 0 5 11 0 .. 0 7 17 1 .. 0 # 20.946 33.479 46.421 # LEU16:C@CG
Env_1aut_6  0 1 0 .. 0 0 1 0 .. 0 1 3 0 .. 0 5 9 0 .. 0 6 10 0 .. 0 6 17 0 .. 0 # 19.619 33.623 47.263 # LEU16:C@CD1
Env_1aut_7  0 1 0 .. 0 0 2 0 .. 0 1 3 0 .. 0 5 8 0 .. 0 5 10 1 .. 0 7 12 2 .. 0 # 22.367 33.517 47.911 # LEU16:C@CD2
Env_1aut_8  0 0 0 .. 0 0 1 0 .. 0 3 3 0 .. 0 3 7 0 .. 0 5 8 1 .. 0 2 14 2 .. 0 # 18.85 37.106 45.557 # LEU16:C@H
Env_1aut_9  0 0 0 .. 0 2 3 0 .. 0 0 3 0 .. 0 3 6 0 .. 0 3 11 0 .. 0 8 15 0 .. 0 # 18.864 34.494 52.085 # ILE17:C@N
Env_1aut_10 0 1 0 .. 0 2 2 0 .. 0 0 2 0 .. 0 2 4 0 .. 0 6 13 0 .. 0 5 13 0 .. 0 # 18.851 33.673 53.533 # ILE17:C@CA
```

Figure B.
Computational representation of a microenvironment. The number of columns per shell have been abbreviated.
Feature Vector files look exactly like this.

The *featurize* program takes points in a biomolecule to build microenvironments (Figure A) and produce computational representations of said microenvironments (Figure B).

The *featurize* program analyzes mutually exclusive concentric spherical volumes (called shells) around a given point in a biomolecule. These shells collectively describe a microenvironment. The *featurize* program tallies physicochemical properties for each atom contained in each shell. These tallies form the computational representation describing the microenvironment. Groups of microenvironment computational representations are called Feature Vectors and are stored in Feature Vector Files.

Machine Learning can be trained on Feature Vectors to produce biomolecule functional class models. Biomolecules of unknown function can be characterized as Feature Vectors and scored against functional class models to predict functionality.

Figure 1: FEATURE Project Details

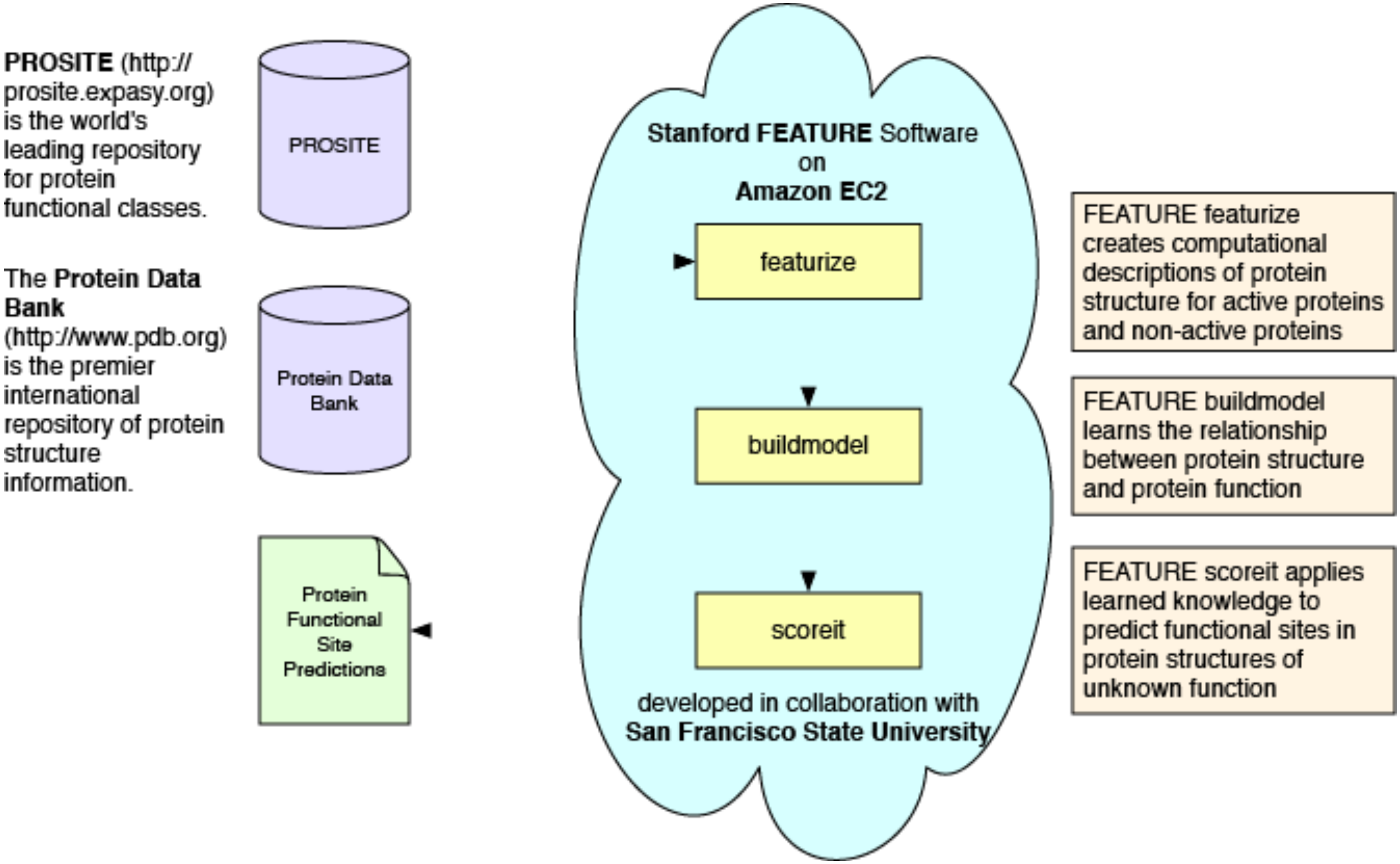
The Challenge

FEATURE, like other innovative scientific projects, has an insatiable appetite for high performance computing and the project's research scientists found that the computational demand for exploring detailed aspects of biological molecules soon outgrew the university's facilities. Computing resources are shared at San Francisco State University and high demand meant that researchers had to re-shape the size and scope of their questions or face long delays for available resources. In addition, these constraints led to long waits for results and put an arbitrary cap on the experiments that the scientists could run.

Why Amazon Web Services

The scientists only needed computational resources periodically and it wasn't cost-effective to purchase a large-scale resource and maintain it for irregular use. As the research team considered their options, they realized that the on-demand access to computational resources provided by Amazon Web Services (AWS) met their purposes. "The pay-as-you-go model of Amazon Elastic Compute Cloud (Amazon EC2) was the most appropriate option versus owning a large server in-house," says Professor Petkovic.

The research team built FEATURE using C, C++, Perl and Python, among other tools. They deployed the cluster to Amazon EC2 with MIT StarCluster, an automated provisioning tool built for scientific and technical high performance computing. The Protein Databank and protein structure databases were loaded onto Amazon Elastic Block Store (Amazon EBS) volumes for easy management and re-use, and are accessed using an Amazon Linux custom machine image (Amazon Linux AMI). Figure 2 demonstrates the architecture of the FEATURE project.


PROSITE
Protein Data Bank
Protein Functional Site Predictions
Stanford FEATURE Software on Amazon EC2
featurize
buildmodel
scoreit
developed in collaboration with San Francisco State University

PROSITE (http://prosite.expasy.org) is the world's leading repository for protein functional classes.

The **Protein Data Bank** (http://www.pdb.org) is the premier international repository of protein structure information.

FEATURE featurize creates computational descriptions of protein structure for active proteins and non-active proteins

FEATURE buildmodel learns the relationship between protein structure and protein function

FEATURE scoreit applies learned knowledge to predict functional sites in protein structures of unknown function

Figure 2: FEATURE Project Architecture

The Benefits

To evaluate the performance of the FEATURE project on AWS, the team used software profiling and I/O benchmarking to measure performance metrics. Petkovic explains, “The team has a small, 40-node in-house cluster. We compared this to the cloud and found that Amazon EC2 was vastly superior in terms of CPU cycles per cost, as well as providing the ability to scale up when needed. Experiments that used to take us weeks now run overnight. This means that our scientists are always engaged and not waiting for results. AWS greatly reduced our turnaround time for scientific inquiry.”

Professor Petkovic estimates that their computing costs have been reduced by about 20 times. “We estimate that a small, 40-node in-house cluster runs at $ 1.71 per computer unit per hour. In comparison, Amazon EC2 costs us only $0.08 per equivalent elastic computer unit (ECU) per hour,” he explains. In addition, Petkovic and his team are able to use billing alerts and other cost optimization tools that AWS provides to plan and manage the cost of using the service.

“AWS provides on-demand access to high performance resources, which enables us to focus on science, rather than the heavy lifting of maintaining server infrastructure. AWS helps us lift the ceiling on the size and scope of our machine learning experiments,” says Petkovic.