Prediction of DNA and RNA binding sites using machine learning approach

Presented by
Ms. Thitima Panyayai 49460010
Bioinformatics Program, King Mongkut’s University of Technology Thonburi
• Introduction
• Predicting DNA and RNA binding sites
• Application
• Conclusion
Outline

- **Introduction**
- Predicting DNA and RNA binding sites
- **Application**
- **Conclusion**
Gene Expression

- An organism may contain many types of somatic cells, each with distinct shape and function.
- However, they all have the same genome. The genes in a genome do not have any effect on cellular functions until they are “expressed”

http://www.web-books.com/
Gene expression means the production of a protein or a functional RNA from its gene. Several steps are required:

- Transcription
- RNA Processing
- Transport
- Translation Synthesis
Levels of gene expression regulation

Epigenetic mechanisms

Chromatin Structure

Transcription

Post transcription

Translation

Protein regulation

RNA regulation

Plagiarized from Kanda Lertladaluck
Levels of gene expression regulation

Gene regulation:

**Epigenetics**
- DNA methylation
- Chromatin remodeling
- Genomic imprinting

Plagiarized from Kanda Lertladaluck
Levels of gene expression regulation

Gene regulation:

By protein

- Transcription factors
- Effectors: activator, repressor

Plagiarized from Kanda Lertladaluck
Levels of gene expression regulation

Gene regulation:

By RNA

- RNAi: siRNA & miRNA
- RNAa: sRNA & Epigenetics
- Loop formation of intergenic region: Attenuation and RBS sequestration
- Ribozymes
- UTR controlling elements: Riboswitches

Plagiarized from Kanda Lertladaluck
A region on a protein, DNA, or RNA to which specific other molecules and ions form a chemical bond.
Any protein that binds to double- or single stranded DNA.

Sequence-specific DNA-binding proteins generally interact with the major groove of B-DNA, because it exposes more functional groups that identify a base pair.
RNA-binding protein

- Typically cytoplasmic and nuclear proteins that associate with (bind) and facilitate the translation of RNAs.

- Some examples include: translation initiation factors that bind RNA, polyA-binding proteins, other RNA binding proteins that regulate translation, etc.
Why identify the binding site?

Understand the molecular details of specific residue – residue contacts that mediate protein-DNA, protein-RNA recognition.

TO advances in drug discovery.

Machine learning

Is concerned with the design and development of algorithms and techniques that allow computers to "learn".

http://en.wikipedia.org/wiki/Machine_learning

Machine learning

Algorithms

(Statistical) Inference

Data Structures
**Machine learning technique**

- Decision trees
- Classification rules
- Instance-based learning
- Clustering
- Neural Network
- Naïve Bayes Classifier
Outline

- Introduction
- Predicting DNA and RNA binding sites
- Application
- Conclusion
Objective

To identify DNA and RNA binding sites using machine learning approaches.

MQLNSVTTEDTGTYCYCTRGNGDWGSVSTSSFRYQGTTLVSSAKTPPS...

Naïve Bayes Classifier
**Prediction of DNA and RNA binding sites using machine learning approach**

**PROBLEM:** Given the sequence of a protein predict which amino acid participate in protein-RNA or protein-DNA interactions

**APPROACH:** Generate datasets of known complexes from PDB to train & test machine learning algorithms (Naïve Bayes)

**GOAL:** Classify each amino acid in target protein as either interface or non-interface residue
Naïve Bayes Classifier

• Is a simple probabilistic classifier based on applying Bayes' theorem.
Learning Scenario

Data set of proteins

Representation of labeled sequences

Training set

Learning algorithms

Classifier

Test set

Predicted amino acid in target protein as interface residue
Naïve Bayes Classifier

Identifying amino acid residues involved in protein-DNA interactions from sequence

Naïve Bayes Classifier

Data set of sequences proteins

C is predicted to be 1 if the ratio likelihood is greater than 0 and 0 otherwise.

\[
P(c = 1 \mid X = x_1x_2...x_n) \\
\frac{P(c = 0 \mid X = x_1x_2...x_n)}{P(c = 1 \mid X = x_1x_2...x_n)} \\
P(c = 1) \prod_{i=1}^{n} P(x_i \mid c = 1) \\
= \frac{P(c = 0) \prod_{i=1}^{n} P(x_i \mid c = 0)}{P(c = 1) \prod_{i=1}^{n} P(x_i \mid c = 1)} > \theta
\]

Naïve Bayes Classifier

Leave-one-out cross-validation

\[
P(c = 1 \mid X = x_1 x_2 \ldots x_n) = \frac{P(c = 1) \prod_{i=1}^{n} P(x_i \mid c = 1)}{P(c = 0) \prod_{i=1}^{n} P(x_i \mid c = 0)} > \theta
\]

### Performance measures

<table>
<thead>
<tr>
<th>Hypothesized class</th>
<th>True class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>p</td>
</tr>
<tr>
<td></td>
<td>True Positives</td>
</tr>
<tr>
<td>N</td>
<td>n</td>
</tr>
<tr>
<td></td>
<td>False Positive</td>
</tr>
<tr>
<td></td>
<td>False Negative</td>
</tr>
<tr>
<td></td>
<td>True Negative</td>
</tr>
</tbody>
</table>

**True Positives**

- **Sensitivity** = \(\frac{TP}{TP + FN}\)

- **Specificity** = \(\frac{TP}{TP + FP}\)

- **Accuracy** = \(\frac{TP + TN}{N}\)

- **Cohen’s \(\kappa\)** = \(\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}\)

---

<table>
<thead>
<tr>
<th>Performance measure</th>
<th>Sequence-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Identities (ID)</td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td>0.25</td>
</tr>
<tr>
<td>Accuracy(%)</td>
<td>77</td>
</tr>
<tr>
<td>Specificity+(%)</td>
<td>37</td>
</tr>
<tr>
<td>Sensitivity+(%)</td>
<td>43</td>
</tr>
</tbody>
</table>

Predictions in The Context of 3-D Structures

Predicted

Actual

Pit-1, PDB 1au7

TP: 30
FP: 16
TN: 86
FN: 14
CC: 0.51
Accuracy: 79%

Predictions Compared With PROSITE Motifs

>lau7A
Sequence : GMRALEQFANEFKVRRIKLGYTNVEALAAVHGSEFSQTTICRFENLQLSFKNACKLK
Interface : * * *** ******** ** * *** ** **
Prediction: * * * **** * **** * ** *
Motifs : ************* *************

- Predicted binding sites substantially overlap with 34 of the 37 “DNA-binding” PROSITE motifs
- In 52 of the 56 proteins, the predictor identifies at least 20% of the DNA-binding residues
- 28 of the 56 proteins contain no PROSITE motifs that are annotated as “DNA-binding”

## Comparison With Previous Study

<table>
<thead>
<tr>
<th>Method</th>
<th>Naïve Bayes classifier</th>
<th>Ahmad and Sarai method*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation Coefficient</td>
<td>0.26</td>
<td>0.23</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>80</td>
<td>66</td>
</tr>
<tr>
<td>Specificity+ (%)</td>
<td>29</td>
<td>21</td>
</tr>
<tr>
<td>Sensitivity+ (%)</td>
<td>48</td>
<td>68</td>
</tr>
</tbody>
</table>

Ahmad and Sarai (2005) used a PSSM-based neural network classifier to identify interface residues in protein-DNA interactions.

- **PSSM** = Position Specific Scoring Matrices
Are form of artificial intelligence which mimic the learning process of the human brain in order to extract patterns from historical data.

http://www.neurosolutions.com/
Neural Network

\[ y = 1 \quad \text{if} \quad \sum_{i=0}^{n} w_i x_i > 0 \]
\[ y = -1 \quad \text{otherwise} \]

http://www.neurosolutions.com/
- **PSSM = Position Specific Scoring Matrices**

## Comparison of Naïve Bayes classifier with neural network classifier

<table>
<thead>
<tr>
<th>Method</th>
<th>Naïve Bayes classifier</th>
<th>Ahmad and Sarai method*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation Coefficient</td>
<td>0.26</td>
<td>0.23</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>80</td>
<td>66</td>
</tr>
<tr>
<td>Specificity+ (%)</td>
<td>29</td>
<td>21</td>
</tr>
<tr>
<td>Sensitivity+ (%)</td>
<td>48</td>
<td>68</td>
</tr>
</tbody>
</table>

Summary

• A simple sequence-based Naive Bayes classifier predicts interface residues in DNA-binding proteins with 75% accuracy, 37% specificity+, 53% sensitivity+ and correlation coefficient of 0.29

• Predicted binding sites
  – correctly indicate the locations of actual binding sites
  – substantially overlap with known PROSITE motifs

Outline

- Introduction
- Predicting DNA and RNA binding sites
- Application
- Conclusion
Predicted protein and RNA binding sites in REV proteins of HIV-1 and EIAV
Rev - a potential target for novel HIV therapies

- **Rev is essential for lentiviral replication**
  - A small nucleoplasmic shuttling protein (HIV Rev 115 aa; EIAV Rev 165 aa)
- Recognizes a specific binding site on viral RNA
  - Rev Responsive Element (RRE)
- Contains specific domains that mediate nuclear localization, RNA binding and nuclear export
- **Rev's critical role in lentiviral replication makes it an attractive target for antiviral (AIDs) therapy**

Terribilini, M., et al., Identify interaction sites in “recalcitrant” proteins: predicted protein and RNA binding sites in REV proteins of HIV-1 and EIAV agree with experimental data,
Macromolecular interactions mediated by the Rev protein in lentiviruses (HIV & EIAV)

Terribilini, M., et al, Identify interaction sites in “recalcitrant” proteins: predicted protein and RNA binding sites in REV proteins of HIV-1 and EIAV agree with experimental data,
Problem: no high resolution Rev structure!

• Why??
  – Rev aggregates at concentrations needed for NMR or X-ray crystallography

• What about insights from sequence comparisons?
  – Very little sequence similarity among different Rev family members (e.g., EIAV vs HIV <10%)

• But: lentiviral Rev proteins are functionally homologous

Terribilini, M., et al, Identify interaction sites in “recalcitrant” proteins: predicted protein and RNA binding sites in REV proteins of HIV-1 and EIAV agree with experimental data,
Hypothesis: Rev proteins share structural features critical for function

Terribilini, M., et al. Identify interaction sites in “recalcitrant” proteins: predicted protein and RNA binding sites in REV proteins of HIV-1 and EIAV agree with experimental data,
- Predicted protein & RNA Binding Sites in Rev proteins of HIV-1 & EIAV agree with available experimental data.

Outline

• Introduction
• Predicting DNA and RNA binding sites
• Application
• Conclusion
Conclusion

• Computational methods can provide insight into protein-protein & protein-nucleic acid interfaces, even for "recalcitrant" proteins whose structures have not been determined

• Integration of experimental results to improve algorithms will be important
Thank you
Representative Machine Learning Applications in Bioinformatics and Computational Biology

- Gene finding
- Ribosome binding site identification
- Promoter identification
- Prediction of protein structural features
- Protein binding site identification
- Prediction of protein function
- Proteome analysis
- Genetic network inference
- Cancer diagnosis
- Gene annotation
What is Machine Learning?

A program $M$ is said to learn from experience $E$ with respect to some class of tasks $T$ and performance measure $P$ if its performance as measured by $P$ on tasks in $T$ in an environment $Z$ improves with experience $E$.

Example 1

T – cancer diagnosis
E – a set of diagnosed cases
P – accuracy of diagnosis on new cases
Z – noisy measurements, occasionally misdiagnosed training cases
M – a program that runs on a general purpose computer
Machine Learning in Context

- Systems Biology
- Bioinformatics
- Chemoinformatics
- Neuroinformatics
- Epidemiology
- Environmental Informatics
- Agricultural Informatics
- Social Informatics
- Medical Informatics
- Computer Science
- Cognitive Science
- Statistics
- Mathematics
• **Applied Statistics** – applied almost always to small data sets, manually by a statistician sometimes assisted by a computer

• **Data mining** – emphasis on large data sets, computational and memory considerations

• **Machine learning** – emphasis on automating the discovery of regularities from data, characterizing what can be learned and under what conditions, obtaining guarantees regarding

\[ \text{Machine Learning} = (\text{Statistical Inference} + \text{Data Structures} + \text{Algorithms}) \]
Leave-one-out Cross Validation

• Leave-one-out is the degenerate case of K-Fold Cross Validation, where K is chosen as the total number of examples
  – For a dataset with N examples, perform N experiments
  – For each experiment use N-1 examples for training and the remaining example for testing

• As usual, the true error is estimated as the average error rate on test examples

\[
E = \frac{1}{N} \sum_{i=1}^{N} E_i
\]
K-Fold Cross-validation

- Create a K-fold partition of the dataset
  - For each of K experiments, use K-1 folds for training and the remaining one for testing

- K-Fold Cross validation is similar to Random Subsampling
  - The advantage of K-Fold Cross validation is that all the examples in the dataset are eventually used for both training and testing

As before, the true error is estimated as the average error rate

\[ E = \frac{1}{K} \sum_{i=1}^{K} E_i \]
• A Naive Bayes classifier requires significantly less computational effort (a single pass through the training data) to train than a neural network classifier (which requires multiple passes through the training data), making it especially well suited for use with large data sets or in settings that call for incremental update of the classifier as new training data become available.
Instances –
ordered 3-tuples of attribute values corresponding to

- Height (tall, short)
- Hair (dark, blonde, red)
- Eye (blue, brown)

Classes –
+ , −

Training Data

<table>
<thead>
<tr>
<th>Instance</th>
<th>Class label</th>
</tr>
</thead>
<tbody>
<tr>
<td>I₁ (t, d, l)</td>
<td>+</td>
</tr>
<tr>
<td>I₂ (s, d, l)</td>
<td>+</td>
</tr>
<tr>
<td>I₃ (t, b, l)</td>
<td>−</td>
</tr>
<tr>
<td>I₄ (t, r, l)</td>
<td>−</td>
</tr>
<tr>
<td>I₅ (s, b, l)</td>
<td>−</td>
</tr>
<tr>
<td>I₆ (t, b, w)</td>
<td>+</td>
</tr>
<tr>
<td>I₇ (t, d, w)</td>
<td>+</td>
</tr>
<tr>
<td>I₈ (s, b, w)</td>
<td>+</td>
</tr>
</tbody>
</table>

Probabilities to estimate

\[
P(+) = \frac{5}{8} \quad P(Height | c) \quad P(Hair | c) \quad P(Eye | c)
\]

|   | P(Height | c) | t | s |
|---|--------|---|---|
| + |        | 3/5 | 2/5 |
| − |        | 2/3 | 1/3 |

|   | P(Hair | c) | d | b | r |
|---|--------|---|---|---|
| + |        | 3/5 | 2/5 | 0 |
| − |        | 0   | 2/3 | 1/3 |

|   | P(Eye | c) | l | w |
|---|--------|---|---|
| + |        | 2/5 | 3/5 |
| − |        | 1   | 0  |

Classify (Height=t, Hair=b, eye=l)
\[
P(X | +) = \frac{(3/5)(2/5)(2/5)}{(12/125)} = \frac{12}{125}
\]
\[
P(X | −) = \frac{(2/3)(2/3)(1)}{(4/9)}
\]

Classification = ?

Classify (Height=t, Hair=r, eye=w)

Note the problem with zero probabilities

Solution – Use Laplacian estimates