[**https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+%28Splice-junction+Gene+Sequences%29**](https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+(Splice-junction+Gene+Sequences))

|  |  |
| --- | --- |
| **Molecular Biology (Splice-junction Gene Sequences) Data Set**  *Download*: [Data Folder](https://archive.ics.uci.edu/ml/machine-learning-databases/molecular-biology/splice-junction-gene-sequences/), [Data Set Description](https://archive.ics.uci.edu/ml/machine-learning-databases/molecular-biology/splice-junction-gene-sequences/splice.names)  **Abstract**: Primate splice-junction gene sequences (DNA) with associated imperfect domain theory |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Sequential, Domain-Theory | **Number of Instances:** | 3190 | **Area:** | Life |
| **Attribute Characteristics:** | Categorical | **Number of Attributes:** | 61 | **Date Donated** | 1992-01-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | No | **Number of Web Hits:** | 92086 |

**Source:**

Creators:   
  
1. All examples taken from Genbank 64.1 (ftp site: genbank.bio.net)   
2. Categories "ei" and "ie" include every "split-gene" for primates in Genbank 64.1   
3. non-splice examples taken from sequences known not to include a splicing site   
  
Donor:   
  
G. Towell, M. Noordewier, and J. Shavlik,   
{towell,shavlik}@cs.wisc.edu, noordewi **'@'** cs.rutgers.edu

**Data Set Information:**

Problem Description:   
  
Splice junctions are points on a DNA sequence at which `superfluous' DNA is removed during the process of protein creation in higher organisms. The problem posed in this dataset is to recognize, given a sequence of DNA, the boundaries between exons (the parts of the DNA sequence retained after splicing) and introns (the parts of the DNA sequence that are spliced out). This problem consists of two subtasks: recognizing exon/intron boundaries (referred to as EI sites), and recognizing intron/exon boundaries (IE sites). (In the biological community, IE borders are referred to a ``acceptors'' while EI borders are referred to as ``donors''.)   
  
This dataset has been developed to help evaluate a "hybrid" learning algorithm (KBANN) that uses examples to inductively refine preexisting knowledge. Using a "ten-fold cross-validation" methodology on 1000 examples randomly selected from the complete set of 3190, the following error rates were produced by various ML algorithms (all experiments run at the Univ of Wisconsin, sometimes with local implementations of published algorithms).   
  
System -- Neither -- EI -- IE   
---------------------------------------------------   
KBANN -- 4.62 -- 7.56 -- 8.47   
BACKPROP -- 5.29 -- 5.74 -- 10.75   
PEBLS -- 6.86 -- 8.18 -- 7.55   
PERCEPTRON -- 3.99 -- 16.32 -- 17.41   
ID3 -- 8.84 -- 10.58 -- 13.99   
COBWEB -- 11.80 -- 15.04 -- 9.46   
Near. Neighbor -- 31.11 -- 11.65 -- 9.09

**Attribute Information:**

1. One of {n ei ie}, indicating the class.   
2. The instance name.   
3-62. The remaining 60 fields are the sequence, starting at position -30 and ending at position +30. Each of these fields is almost always filled by one of {a, g, t, c}. Other characters indicate ambiguity among the standard characters according to the following table:   
  
character: meaning   
D: A or G or T   
N: A or G or C or T   
S: C or G   
R: A or G

**Relevant Papers:**

M. O. Noordewier and G. G. Towell and J. W. Shavlik, 1991; "Training Knowledge-Based Neural Networks to Recognize Genes in DNA Sequences". Advances in Neural Information Processing Systems, volume 3, Morgan Kaufmann.   
[[Web Link]](http://rexa.info/paper/fa41b2382523a41a7842eac3dda0625f7869897a)   
  
G. G. Towell and J. W. Shavlik and M. W. Craven, 1991; "Constructive Induction in Knowledge-Based Neural Networks", In Proceedings of the Eighth International Machine Learning Workshop, Morgan Kaufmann.   
[[Web Link]](http://rexa.info/paper/a3d216a5362c7604c773adb70488eb93a41ded2d)   
  
G. G. Towell, 1991; "Symbolic Knowledge and Neural Networks: Insertion, Refinement, and Extraction", PhD Thesis, University of Wisconsin - Madison.   
[[Web Link]](http://rexa.info/paper/176aca97d8e0ff00dac888df879d3fa4aed603ca)   
  
G. G. Towell and J. W. Shavlik, 1992; "Interpretation of Artificial Neural Networks: Mapping Knowledge-based Neural Networks into Rules", In Advances in Neural Information Processing Systems, volume 4, Morgan Kaufmann.   
[[Web Link]](http://rexa.info/paper/38d40238d3c994c7d26766af72b2de9e327ee5f5)

**Papers That Cite This Data Set1:**



Jinyan Li and Limsoon Wong. [Using Rules to Analyse Bio-medical Data: A Comparison between C4.5 and PCL](http://rexa.info/paper/4f29b49d0d83edf9e0a21d102467df5ffdf64fcb). WAIM. 2003. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "4f29b49d0d83edf9e0a21d102467df5ffdf64fcb)].  
  
Susanne Hoche and Stefan Wrobel. [Scaling Boosting by Margin-Based Inclusionof Features and Relations](http://rexa.info/paper/f877b9d13dc80d6292789ea42108706658a378ab). ECML. 2002. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "f877b9d13dc80d6292789ea42108706658a378ab)].  
  
S. Sathiya Keerthi and Kaibo Duan and Shirish Krishnaj Shevade and Aun Neow Poo. [A Fast Dual Algorithm for Kernel Logistic Regression](http://rexa.info/paper/6a6b4ca13137132e00185012fd7c4667ff964769). ICML. 2002. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "6a6b4ca13137132e00185012fd7c4667ff964769)].  
  
Michael G. Madden. [Evaluation of the Performance of the Markov Blanket Bayesian Classifier Algorithm](http://rexa.info/paper/d25c86892bfa80849ecd629de4c45936a52a1590). CoRR, csLG/0211003. 2002. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "d25c86892bfa80849ecd629de4c45936a52a1590)].  
  
Xiaojin Zhu. [Label Propagation for Eukaryotic Splice Junction Identification](http://rexa.info/paper/75f0f1770a1317fc5ab730609fa1e3cf642794f6). 2002. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "75f0f1770a1317fc5ab730609fa1e3cf642794f6)].  
  
Mukund Deshpande and George Karypis. [Evaluation of Techniques for Classifying Biological Sequences](http://rexa.info/paper/f006e50cdba5e00538e2b48452e1d0bab8ed02b9). PAKDD. 2002. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "f006e50cdba5e00538e2b48452e1d0bab8ed02b9)].  
  
Jinyan Li and Kotagiri Ramamohanarao and Guozhu Dong. [Combining the Strength of Pattern Frequency and Distance for Classification](http://rexa.info/paper/0cedee7ef9325127fb946dcdddf8795644e28335). PAKDD. 2001. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "0cedee7ef9325127fb946dcdddf8795644e28335)].  
  
Jinyan Li and Guozhu Dong and Kotagiri Ramamohanarao and Limsoon Wong. [DeEPs: A New Instance-based Discovery and Classification System](http://rexa.info/paper/ca1c2b19089d0ca0d4069b04c35d6f6c4312db79). Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases. 2001. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "ca1c2b19089d0ca0d4069b04c35d6f6c4312db79)].  
  
Jinyan Li and Guozhu Dong and Kotagiri Ramamohanarao. [Instance-Based Classification by Emerging Patterns](http://rexa.info/paper/111c890abe7dc9f26523d1c59b7a75a04b69256b). PKDD. 2000. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "111c890abe7dc9f26523d1c59b7a75a04b69256b)].  
  
Thomas G. Dietterich. [An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization](http://rexa.info/paper/4eff6dda32898a8bdc8268d2a790a96b93f2e262). Machine Learning, 40. 2000. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "4eff6dda32898a8bdc8268d2a790a96b93f2e262)].  
  
Marina Meila and Michael I. Jordan. [Learning with Mixtures of Trees](http://rexa.info/paper/4a338bab101f6721ea5dc4d76de5456721fbf0a3). Journal of Machine Learning Research, 1. 2000. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "4a338bab101f6721ea5dc4d76de5456721fbf0a3)].  
  
Kai Ming Ting and Ian H. Witten. [Issues in Stacked Generalization](http://rexa.info/paper/70172e511a3bc27c7927119a3b2a3405fbad99e0). J. Artif. Intell. Res. (JAIR, 10. 1999. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "70172e511a3bc27c7927119a3b2a3405fbad99e0)].  
  
Yoav Freund and Lorne Mason. [The Alternating Decision Tree Learning Algorithm](http://rexa.info/paper/58642563848aaece947320829da7d338224cb8b3). ICML. 1999. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "58642563848aaece947320829da7d338224cb8b3)].  
  
Lorne Mason and Jonathan Baxter and Peter L. Bartlett and Marcus Frean. [Boosting Algorithms as Gradient Descent](http://rexa.info/paper/f820b10d6723504d343c9741f7333fbc3393fd05). NIPS. 1999. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "f820b10d6723504d343c9741f7333fbc3393fd05)].  
  
Kagan Tumer and Nikunj C. Oza. [Decimated Input Ensembles for Improved Generalization](http://rexa.info/paper/2b8c2489b0b9622471466385b6d5c1cf703bd4c8). NASA Ames Research Center. 1999. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "2b8c2489b0b9622471466385b6d5c1cf703bd4c8)].  
  
Blaz Zupan and Marko Bohanec and Janez Dem#sar and Ivan Bratko. [Learning by Discovering Concept Hierarchies](http://rexa.info/paper/50aaf7ca6a34bd2aadbd1cb87235bda27650bb18). Artif. Intell, 109. 1999. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "50aaf7ca6a34bd2aadbd1cb87235bda27650bb18)].  
  
Foster J. Provost and Tom Fawcett and Ron Kohavi. [The Case against Accuracy Estimation for Comparing Induction Algorithms](http://rexa.info/paper/3ce34105eca9577e44de969f9ef94ca3cdcad3d9). ICML. 1998. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "3ce34105eca9577e44de969f9ef94ca3cdcad3d9)].  
  
Andreas L. Prodromidis. [On the Management of Distributed Learning Agents Ph.D. Thesis Proposal CUCS-032-97](http://rexa.info/paper/f23deedb8c4f89030e936b0d119459b41278db96). Department of Computer Science Columbia University. 1998. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "f23deedb8c4f89030e936b0d119459b41278db96)].  
  
Manoranjan Dash and Huan Liu. [Hybrid Search of Feature Subsets](http://rexa.info/paper/8fe250a2bb9e7ca099dfd176a16a08f24bb95939). PRICAI. 1998. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "8fe250a2bb9e7ca099dfd176a16a08f24bb95939)].  
  
Adam J. Grove and Dale Schuurmans. [Boosting in the Limit: Maximizing the Margin of Learned Ensembles](http://rexa.info/paper/c334718c7e09d2671197a8c9526018de1d816903). AAAI/IAAI. 1998. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "c334718c7e09d2671197a8c9526018de1d816903)].  
  
Kai Ming Ting and Boon Toh Low. [Model Combination in the Multiple-Data-Batches Scenario](http://rexa.info/paper/6a9871da32f3042a5de9082a36a11d5aecda6df8). ECML. 1997. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "6a9871da32f3042a5de9082a36a11d5aecda6df8)].  
  
Kamal Ali and Michael J. Pazzani. [Error Reduction through Learning Multiple Descriptions](http://rexa.info/paper/349644828ad3c8da8e1a15eee5aec0d704db34a1). Machine Learning, 24. 1996. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "349644828ad3c8da8e1a15eee5aec0d704db34a1)].  
  
Pedro Domingos. [Using Partitioning to Speed Up Specific-to-General Rule Induction](http://rexa.info/paper/3b8e310b092f79f90f6c4aac5dda1185d5a1aac7). Department of Information and Computer Science University of California, Irvine. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "3b8e310b092f79f90f6c4aac5dda1185d5a1aac7)].  
  
Kai Ming Ting and Ian H. Witten. [Stacked Generalization: when does it work](http://rexa.info/paper/939595ca638eb3390e9bb9c4e6cc1352163cbf18). Department of Computer Science University of Waikato. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "939595ca638eb3390e9bb9c4e6cc1352163cbf18)].  
  
Cesar Guerra-Salcedo and Stephen Chen and Darrell Whitley and Sarah Smith. [Fast and Accurate Feature Selection Using Hybrid Genetic Strategies](http://rexa.info/paper/379f34d22ceaf54a97c9a95f0cb0e6eb2a44c1dc). Department of Computer Science Colorado State University. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "379f34d22ceaf54a97c9a95f0cb0e6eb2a44c1dc)].  
  
Gustavo E. A and Gustavo E A P A Batista and Ronaldo C. Prati and Maria Carolina Monard. [A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data](http://rexa.info/paper/a27103950b8d7098468bd52f2af1d033fed91419). Instituto de Ci ^ encias Matem aticas e de Computac~ ao. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "a27103950b8d7098468bd52f2af1d033fed91419)].  
  
Kai Ming Ting and Boon Toh Low. [Theory Combination: an alternative to Data Combination](http://rexa.info/paper/49c65a0dfe258129414e27f6a3b7c42b7fbe426e). University of Waikato. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "49c65a0dfe258129414e27f6a3b7c42b7fbe426e)].  
  
Rudy Setiono. [Extracting M-of-N Rules from Trained Neural Networks](http://rexa.info/paper/42ba5137ab9d88dc1eae5caac18cb6a1818ae700). School of Computing National University of Singapore. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "42ba5137ab9d88dc1eae5caac18cb6a1818ae700)].  
  
Rong-En Fan and P. -H Chen and C. -J Lin. [Working Set Selection Using the Second Order Information for Training SVM](http://rexa.info/paper/da329267bf8880c2becb15eae121a5b002347349). Department of Computer Science and Information Engineering National Taiwan University. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "da329267bf8880c2becb15eae121a5b002347349)].  
  
M. A. Galway and Michael G. Madden. [DEPARTMENT OF INFORMATION TECHNOLOGY technical report NUIG-IT-011002 Evaluation of the Performance of the Markov Blanket Bayesian Classifier Algorithm](http://rexa.info/paper/d342517262ff52ffd3566bd8f520b36723486aa3). Department of Information Technology National University of Ireland, Galway. [[View Context](https://archive.ics.uci.edu/ml/support/Molecular+Biology+(Splice-junction+Gene+Sequences)" \l "d342517262ff52ffd3566bd8f520b36723486aa3)].

**Citation Request:**

Please refer to the Machine Learning Repository's [citation policy](https://archive.ics.uci.edu/ml/citation_policy.html)