Package ‘EncDNA’

October 12, 2022

Type Package

Title Encoding of Nucleotide Sequences into Numeric Feature Vectors

Version 1.0.2

Date 2019-05-28

Author Prabina Kumar Meher

Maintainer Prabina Kumar Meher <meherprabin@yahoo.com>

Depends R(>= 3.3.0)

biocViews

Imports Biostrings

LazyData TRUE

Description We describe fifteen different splice site sequence encoding schemes that have been used in earlier studies for mapping of splice site sequences into numeric feature vectors. These encoding schemes will also be helpful for transforming other nucleotide sequences into numeric forms, provided they are of equal length. These encoding schemes will help the computational biologist working in the field of classification (binary or multiclass) or prediction involving nucleic acid sequences of equal length.

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2019-05-28 06:00:03 UTC

R topics documented:

`APR.Feature` *Adjacent position relationship feature.*

Description

This feature was proposed by Li *et al*.(2012). In fact this is similar to the PN.FDTF encoding scheme (Huang *et al*., 2006). In this encoding, correlation between adjacent nucleotides are taken into account. For any nucleotide sequence with n nucleotides, every two consecutive positions between 1 and n, i.e., $(1, 2), (2, 3)...(n - 1, n)$ constitute an APR feature set. For each pair of positions, frequencies of 16 dinucleotides are first computed for both positive and negative dataset, and then the diffence matrix is obtained by substracting the $16 * (n - 1)$ dinucleotide frequency matrix of positive set from that of negative set. The difference matrix is then be used for encoding of nucleotide sequences. In this encoding procedure each sequence with n nucleotides can be encoded into a vector of $(n - 1)$ numeric observations.

Usage

```
APR.Feature(positive_class, negative_class, test_seq)
```

Arguments

`positive_class`	Nucleotide sequence dataset of positive class, must be an object of class `DNAStringSet`.
`negative_class`	Nucleotide sequence dataset of negative class, must be an object of class `DNAStringSet`.
`test_seq`	Nucleotide sequences to be encoded into numeric feature vectors, must be an object of class `DNAStringSet`.

Details

The class DNAStringSet can be obtained by using the function *readDNAStringSet* avialble in **Biostrings** package Bioconductor. Here, the sequences must be supplied in FASTA format. Both positive and negative datasets are required for this encoding scheme.

Value

A numeric matrix of order $m * (n - 1)$, where m is the number of sequences in `test_seq` and n is the length of sequence.

Author(s)

Prabina Kumar Meher, Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

References

Li, J.L., Wang, L.F., Wang, H.Y., Bai, L.Y. and Yuan, Z.M. (2012). High-accuracy splice sites prediction based on sequence component and position features. *Genetics and Molecular Research*, 11(3): 3432-3451.

See Also

`PN.Fdtf.Feature`, `WAM.Feature`

Examples

```
data(droso)
positive <- droso$positive
negative <- droso$negative
test <- droso$test
pos <- positive[1:200]
neg <- negative[1:200]
tst <- test
enc <- APR.Feature(positive_class=pos, negative_class=neg, test_seq=tst)
enc
```

Bayes.Feature *Projecting nucleotide sequences into numeric feature vectors using Bayes kernel encoding approach.*

Description

This sequence encoding technique was introduced by Zhang *et al.* (2006) for prediction of splice sites. In this encoding technique, positional frequencies of nucleotides are computed for both positive and negative datasets, which are then used for encoding of any nucleotide sequence of same length. Each sequence of length *L* can be encoded into a numeric feature vector of length *2L*. Both positive and negative classes of sequences are required for sequence encoding.

Usage

```
Bayes.Feature(positive_class, negative_class, test_seq)
```

Arguments

`positive_class`	Nucleotide sequence dataset of positive class, must be an object of class `DNAStringSet`.
`negative_class`	Nucleotide sequence dataset of negative class, must be an object of class `DNAStringSet`.
`test_seq`	Nucleotide sequences to be encoded into numeric feature vectors, must be an object of class `DNAStringSet`.

Details

The class DNAStringSet can be obtained by using the function *readDNAStringSet* avialble in **Biostrings** package of Bioconductor. Here, the sequences must be supplied in FASTA format.

Value

A numeric matrix of order $m * 2n$, where m is the number of sequences in `test_seq` and n is the sequence length.

Author(s)

Prabina Kumar Meher, Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

References

Zhang, Y., Chu, C., Chen, Y., Zha, H. and Ji, X. (2006). Splice site prediction using support vector machines with a Bayes kernel. *Expert Systems with Applications*, 30: 73-81.

See Also

`POS.Feature`, `MN.Fdtf.Feature`

Examples

```
data(droso)
positive <- droso$positive
negative <- droso$negative
test <- droso$test
pos <- positive[1:200]
neg <- negative[1:200]
tst <- test
enc <- Bayes.Feature(positive_class=pos, negative_class=neg, test_seq=tst)
enc
```

`Density.Feature` *Nucleotide sequence encoding with the distribution of trinucleotides.*

Description

Each nucleotide sequence is encoded into a numeric vector of same length based on the distribution of nucleotides over the sequence. Here, two classes of dataset are not required for encoding, and each sequence is independently encoded instead. This encoding seheme was introduced by Wei *et al*. (2013) for prediction of donor and acceptor human splice sites along with the `MM1.Feature`.

Usage

```
Density.Feature(test_seq)
```

Arguments

`test_seq` Sequence dataset to be encoded, must be an object of class `DNAStringSet`.

Details

The class DNAStringSet can be obtained by reading FASTA sequences using the function *readDNAStringSet* avialble in **Biostrings** package of Bioconductor.

Value

A numeric matrix of order $m * n$, where m is the number of sequences in `test_seq` and n is the length of sequence.

Author(s)

Prabina Kumar Meher, Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

References

Bari, A.T.M.G., Reaz, M.R. and Jeong, B.S. (2014). Effective DNA encoding for splice site prediction using SVM. *MATCH Commun. Math. Comput. Chem.*, 71: 241-258.

Examples

```
data(droso)
test <- droso$test
tst <- test[1:5]
enc <- Density.Feature(test_seq=tst)
enc
```

`droso` *An example dataset consisting of true and false donor splice sites of Drosophila melanogaster.*

Description

The dataset contains 400 true donor sites, 400 false donor sites and 50 false sites as test set. Each sequence is of 40 nucleotides long. The conserved di-nucleotides GT are present at 21^st and 22^nd positions respectively.

Usage

```
data("droso")
```

Format

The dataset belongs to the class `DNAStringSet`.

Source

This is a sample dataset which was collected from the dataset originally developed by Reese *et al*. (1997). The dataset can also be accessed at http://www.fruitfly.org/sequence/drosophila-datasets.html.

References

Reese, M. G., Eeckman, F. H., Kulp, D. and Haussler, D. (1997). Improved splice site detection in Genie. *Journal of Computational Biology*, 4(3): 311-323.

Examples

```
data(droso)
```

Maldoss.Feature *Encoding of nucleic acid sequences using di-nucleotide frequency difference between positive and negative class datasets.*

Description

In Maldoss (Meher *et al*., 2016), the authors propose three encoding approaches namely P1, P2 and P3. Out of these three encoding schemes, the accuracies were reported to be higher for P1 as compared to the other two encoding procedures. Here, we describe the sequence encodng based on P1 only. This P1 encoding approach has similarity with that of PN-FDTF encoding (Huang *et al*., 2006) approach. The difference is only with respect to the logarithmic transformation in case of `Maldoss.Feature`. In this encoding procedure, both positive and negative class sequences are required for transformation of nucleotide sequences into numeric vectors.

Usage

```
Maldoss.Feature(positive_class, negative_class, test_seq)
```

Arguments

`positive_class`	Sequence dataset of the positive class, must be an object of class `DNAStringSet`.
`negative_class`	Sequence dataset of the negative class, must be an object of class `DNAStringSet`.
`test_seq`	Sequences to be encoded into numeric vectors, must be an object of class `DNAStringSet`.

Details

For getting an object of class `DNAStringSet`, the FASTA sequence dataset must be read in R through the function *raedDNAStringSet* available in **Biostrings** package of Bioconductor (https://bioconductor.org/packages/release/bioc/html/Biostrings.html).

Value

A numeric matrix of order $m * (n - 1)$, where m is the number of sequences in `test_seq` and n is the length of sequence.

Author(s)

Prabina Kumar Meher, Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

References

1. Meher, P.K., Sahu, T.K. and Rao, A.R. (2016). Prediction of donor splice sites using random forest with a new sequence encoding approach. *BioData Mining*, 9.
2. Huang, J., Li, T., Chen, K. and Wu, J. (2006). An approach of encoding for prediction of splice sites using SVM. *Biochimie*, 88(7): 923-929.

See Also

`PN.Fdtf.Feature`, `MM1.Feature`, `WAM.Feature`

Examples

```
data(droso)
positive <- droso$positive
negative <- droso$negative
test <- droso$test
pos <- positive[1:200]
neg <- negative[1:200]
tst <- test
enc <- Maldoss.Feature(positive_class=pos, negative_class=neg, test_seq=tst)
enc
```

`MM1.Feature` *Transforming nucleotide sequences into numeric vectors using first order nucleotide dependency.*

Description

The concept of sequence encoding using Markov model (1^{st} order) was introduced by Ho and Rajapakse (2005) for prediction of splice sites. However, this encoding scheme has been comprehensively used by Baten *et al.* (2006) for prediction of splice sites. In this encoding procedure, first order dependencies between nucleotides in nucleotide sequence are accounted. Only the postive class dataset is used for estimation of dependencies in terms of probabilities, which are then used for encoding.

Usage

```
MM1.Feature(positive_class, test_seq)
```

Arguments

`positive_class` Sequence dataset of the positive class, must be an object of class `DNAStringSet`.

`test_seq` Sequences to be encoded into numeric vectors, must be an object of class `DNAStringSet`.

Details

The FASTA sequences should be read into R using the function *readDNAStringSet* available in **Biostrings** package. This encoding is similar to PN.FDTF feature, as far as the dependency among nucleotides in a sequence is concerned. The only difference is the use of positive class only in stead of both positive and negative classes in PN.FDTF. This encoding approach has similarity with WAM features (Meher *et al.* 2016) in which the dinucleotide dependencies are considered.

Value

A numeric matrix of order $m * (n - 1)$, where m is the number of sequences in `test_seq` and n is the length of sequence.

Author(s)

Prabina Kumar Meher, Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

References

1. Rajapakse, J. and Ho, L.S. (2005). Markov encoding for detecting signals in genomic sequences. *IEEE/ACM Trans Comput Biol Bioinf.*, 2(2): 131-142.
2. Baten, A., Chang, B., Halgamuge, S. and Li, J. (2006) Splice site identification using probabilistic parameters and SVM classification. *BMC Bioinformatics*, 7(Suppl 5): S15.
3. Meher, P.K., Sahu, T.K., Rao, A.R. and Wahi, S.D. (2016). Identification of donor splice sites using support vector machine: a computational approach based on positional, compositional and dependency features. *Algorithms for Molecular Biology*, 11(1), 16.

See Also

`PN.Fdtf.Feature`, `WAM.Feature`

Examples

```
data(droso)
positive <- droso$positive
test <- droso$test
pos <- positive[1:200]
tst <- test
enc <- MM1.Feature(positive_class=pos, test_seq=tst)
enc
```

`MM2.Feature` *Mapping nucleotide sequences onto numeric feature vectors based on second order nucleotide dependencies.*

Description

This encoding procedure is similar to the MM1 encoding. The only difference is consideration of second order dependencies unlike first order in `MM2.Feature`. This technique was first conceptualized by Rajapakse and Ho (2005), and adopted by Maji and Garg (2014). The number of parameters to be estimated in MM2 is 64, which is higher than that of MM1 i.e., 16. Further, only the positive class dataset is used for encoding of sequences.

Usage

```
MM2.Feature(positive_class, test_seq)
```

Arguments

`positive_class`	Sequence dataset of the positive class, must be an object of class `DNAStringSet`.
`test_seq`	Sequences to be encoded into numeric vectors, must be an object of class `DNAStringSet`.

Details

For getting an object of class `DNAStringSet`, the FASTA sequences should be read using the function *readDNAStringSet* avialble in the **Biostrings** package.

Value

A numeric matrix of order $m * (n - 2)$, where m is the number of sequences in `test_seq` and n is the length of sequence.

Author(s)

Prabina Kumar Meher, Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

References

1. Rajapakse, J. and Ho, L.S. (2005). Markov encoding for detecting signals in genomic sequences. *IEEE/ACM Trans Comput Biol Bioinf.*, 2(2): 131-142.
2. Maji, S. and Garg, D. (2014). Hybrid approach using SVM and MM2 in splice site junction identification. *Current Bioinformatics*, 9(1): 76-85.

See Also

`MM1.Feature`, `WAM.Feature`

Examples

```
data(droso)
positive <- droso$positive
test <- droso$test
pos <- positive[1:200]
tst <- test
enc <- MM2.Feature(positive_class=pos, test_seq=tst)
enc
```

`MN.Fdtf.Feature` *Sequence encoding with nucleotide frequency difference between two classes of sequence datasets.*

Description

In this encoding procedure, at first, frequency of each nucleotide at each position is computed for both positive and negative classes datasets. Then, the frequency matrix of the positive set is substracted from that of negative set. The sequences are then encoded into numeric vectors after passing them through this difference matrix. So, both positive and negative datasets are necessary for encoding of sequences. This concept was introduced by Huang *et al.* (2006), and was also used by Pashaei *et al.* (2016) to generate features for prediction of splice sites along with other features. This has similarity with Bayes kernel encoding (Zhang *et al.*, 2006), where both frequency matrices are used for encoding instead of the difference matrix.

Usage

```
MN.Fdtf.Feature(positive_class, negative_class, test_seq)
```

Arguments

`positive_class`	Sequence dataset of the positive class, must be an object of class `DNAStringSet`.
`negative_class`	Sequence dataset of the negative class, must be an object of class `DNAStringSet`.
`test_seq`	Sequences to be encoded into numeric vectors, must be an object of class `DNAStringSet`.

Details

For getting an object of class `DNAStringSet`, the sequence dataset must be read in FASTA format through the function *readDNAStringSet* available in the **Biostrings** package of Bioconductor (`https://bioconductor.org/packages/release/bioc/html/Biostrings.html`).

Value

A numeric matrix of order $m * n$, where m is the number of sequences in `test_seq` and n is the sequence length.

Note

This feature does not take into consideration the dependencies among nucleotides in the sequence.

Author(s)

Prabina Kumar Meher, Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

References

1. Zhang, Y., Chu, C., Chen, Y., Zha, H. and Ji, X. (2006). Splice site prediction using support vector machines with a Bayes kernel. *Expert Systems with Applications*, 30: 73-81.
2. Huang, J., Li, T., Chen, K. and Wu, J. (2006). An approach of encoding for prediction of splice sites using SVM. *Biochimie*, 88(7): 923-929.
3. Pashaei, E., Yilmaz, A., Ozen, M. and Aydin, N. (2016). Prediction of splice site using AdaBoost with a new sequence encoding approach. *In Systems, Man, and Cybernetics (SMC), IEEE International Conference*, pp 3853-3858.

See Also

`WMM.Feature`, `Bayes.Feature`, `PN.Fdtf.Feature`

Examples

```
data(droso)
positive <- droso$positive
negative <- droso$negative
test <- droso$test
pos <- positive[1:200]
neg <- negative[1:200]
tst <- test
enc <- MN.Fdtf.Feature(positive_class=pos, negative_class=neg, test_seq=tst)
enc
```

`PN.Fdtf.Feature` *Conversion of nucleotide sequences into numeric feature vectors based on the difference of dinucleotide frequency.*

Description

Dinucleotide frequency matrix is first computed for both positive and negative classes. Then, frequency matrix of the positive class is substracted from that of negative class. The sequences are then passed through this difference matrix to encode them into numeric feature vectors. Similar to the `MN.Fdtf` feature, both positive and negative classes are necessary for encoding of nucleotide sequences. This was also conceptualized by Huang *et al.* (2006). This has also been used by Pashaei *et al.* (2016) as one of the features for prediction of splice sites along with the other features.

Usage

```
PN.Fdtf.Feature(positive_class, negative_class, test_seq)
```

Arguments

`positive_class` Sequence dataset of the positive class, must be an object of class `DNAStringSet`.

`negative_class` Sequence dataset of the negative class, must be an object of class `DNAStringSet`.

`test_seq` Sequences to be encoded into numeric vectors, must be an object of class `DNAStringSet`.

Details

For getting an object of class DNAStringSet, the sequence dataset must be read in FASTA format through the function *readDNAStringSet* available in **Biostrings** package of Bioconductor (https://bioconductor.org/packages/release/bioc/html/Biostrings.html).

Value

A numeric matrix of order $m * (n - 1)$, where m is the number of sequences in test_seq and n is the sequence length.

Note

Both positive and negative classes datasets are essential for the encoding. This feature has similarity with that of MM1.Feature and WAM.Feature with respect to the first order dependency. Unlike MN.Fdtf.Feature, this feature takes into account the first order dependencies of nucleotides in the sequence.

Author(s)

Prabina Kumar Meher, Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

References

1. Huang, J., Li, T., Chen, K. and Wu, J. (2006). An approach of encoding for prediction of splice sites using SVM. *Biochimie*, 88(7): 923-929.
2. Pashaei, E., Yilmaz, A., Ozen, M. and Aydin, N. (2016). Prediction of splice site using AdaBoost with a new sequence encoding approach. *In Systems, Man, and Cybernetics (SMC), IEEE International Conference*, pp 3853-3858.

See Also

MN.Fdtf.Feature, WAM.Feature, MM1.Feature,

Examples

```
data(droso)
positive <- droso$positive
negative <- droso$negative
test <- droso$test
pos <- positive[1:200]
neg <- negative[1:200]
tst <- test
enc <- PN.Fdtf.Feature(positive_class=pos, negative_class=neg, test_seq=tst)
enc
```

POS.Feature *Transformation of nucleic acid sequences into numeric vectors using position-wise frequency of nucleotides.*

Description

This encoding scheme was devised by Li *et al*. (2012). Frequencies of 4 nucleotides are first computed at each position for both positive and negative datasets, resulting in two $4 * L$ probability tables for the two classes for sequence length L. A $4 * L$ statistical difference table is obtained by elementwise substraction of the two probability distribution tables, which is then used for encoding of sequences. Further, as per sparse encoding, the nucleotides A, T, G and C can be encoded as (1,0,0,0), (0,1,0,0), (0,0,1,0) and (0,0,0,1) respectively. The value 1 of sparse encoding is then replaced with the difference values obtained from the difference table for encoding nucleotide at each postion. Thus, it can be said that POS feature encoding is a blending of MN-FDTF (Huang *et al*., 2006) and Sparse encoding (Meher *et al*., 2016) technique.

Usage

```
POS.Feature(positive_class, negative_class, test_seq)
```

Arguments

`positive_class`	Sequence dataset of the positive class, must be an object of class `DNAStringSet`.
`negative_class`	Sequence dataset of the negative class, must be an object of class `DNAStringSet`.
`test_seq`	Sequences to be encoded into numeric vectors, must be an object of class `DNAStringSet`.

Details

The `DNAstringSet` object can be obtained by reading the sequences in FASTA format using the function *readDNAStringSet*available in the **Biostrings** package of Bioconductor.

Value

A numeric matrix of order $m * 4n$, where m is the number of sequences in `test_seq` and n is the length of sequence.

Note

In this encoding procedure, dependencies of nucleotides are not taken into consideration. Both positive and negative datasets are required for encoding of nucleotide sequences. Each sequence of length L can be transformed into a numeric vector of length $4 * L$ with this encoding technique.

Author(s)

Prabina Kumar Meher, Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

References

1. Huang, J., Li, T., Chen, K. and Wu, J. (2006). An approach of encoding for prediction of splice sites using SVM. *Biochimie*, 88(7): 923-929.
2. Li, J.L., Wang, L.F., Wang, H.Y., Bai, L.Y., Yuan, Z.M. (2012). High-accuracy splice sites prediction based on sequence component and position features. *Genetics and Molecular Research*, 11(3): 3432-3451.
3. Meher, P.K., Sahu, T.K., Rao, A.R. and Wahi, S.D. (2016). A computational approach for prediction of donor splice sites with improved accuracy. *Journal of Theoretical Biology*, 404: 285-294.

See Also

`MN.Fdtf.Feature`, `Bayes.Feature`, `WMM.Feature`

Examples

```
data(droso)
positive <- droso$positive
negative <- droso$negative
test <- droso$test
pos <- positive[1:200]
neg <- negative[1:200]
tst <- test
enc <- POS.Feature(positive_class=pos, negative_class=neg, test_seq=tst)
enc
```

`Predoss.Feature` *Encoding nucleotide sequences using all possible di-nucleotide dependencies.*

Description

In this encoding, not only the adjecent dependencies are considered, but also the association that exists among non-adjacent nucleotides. In MM1, PN.FDTF features, only the dependencies between adjacent nucleotides are taken into account. Though all possible pair-wise dependencies are first introduced by Meher *et al.* (2014) for predicting splice sites through probablistic approach, the same authors further used this association to encode the splice site dataset for prediction using machine learning classifiers (Meher *et al.*, 2016).

Usage

```
Predoss.Feature(positive_class, negative_class, test_seq)
```

Arguments

`positive_class`	Sequence dataset of the positive class, must be an object of class `DNAStringSet`.
`negative_class`	Sequence dataset of the negative class, must be an object of class `DNAStringSet`.
`test_seq`	Sequences to be encoded into numeric vectors, must be of an object of class `DNAStringSet`.

Details

This encoding approach will be helpful for transformation of nucleotide sequences into numeric feature vectors, which can subsequently be used as input in several supervised learning models for classification.

Value

A numeric matrix of order $m * n^2$, where m is the number of sequences in `test_seq` and n is the length of sequence.

Note

Dimension of the feature space will increase geometrically with increase in the length of the sequence.

Author(s)

Prabina Kumar Meher, Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

References

1. Meher, P.K., Sahu, T.K., Rao, A.R. and Wahi, S.D. (2014). A statistical approach for 5' splice site prediction using short sequence motifs and without encoding sequence data. *BMC Bioinformatics*, 15(1), 362.
2. Meher, P.K., Sahu, T.K., Rao, A.R. and Wahi, S.D. (2016). A computational approach for prediction of donor splice sites with improved accuracy. *Journal of Theoretical Biology*, 404: 285-294.

Examples

```
data(droso)
positive <- droso$positive
negative <- droso$negative
test <- droso$test
pos <- positive[1:200]
neg <- negative[1:200]
tst <- test
enc <- Predoss.Feature(positive_class=pos, negative_class=neg, test_seq=tst)
enc
```

`SAE.Feature` *Encoding of nucleotide sequences based on sum of absolute error (SAE) of each sequence.*

Description

The sum of absolute error (SAE) concept was introduced by Meher *et al*. (2014) for prediction of donor splice sites, and was subsequently used by the same authors (Meher *et al*., 2016) for encoding of splice site motif for prediction using supervised learning model. In this encoding technique also all possible pair-wise nucleotide dependencies are considered.

Usage

```
SAE.Feature(positive_class, negative_class, test_seq)
```

Arguments

`positive_class`	Sequence dataset of the positive class, must be an object of class `DNAStringSet`.
`negative_class`	Sequence dataset of the negative class, must be an object of class `DNAStringSet`.
`test_seq`	Sequences to be encoded into numeric vectors, must be an object of class `DNAStringSet`.

Details

In this encoding approach a vector of two observations will be obtained for each sequence. This two values correspond to the values obtained, when only positive class and both positive & neagtive datasets are used for encoding. This encoding scheme is invariant to the length of the sequence. Thus, both positive and negative classes datasets are required for encoding of sequence.

Value

A numeric matrix of order $m * 2$, where m is the number of sequences in `test_seq`.

Author(s)

Prabina Kumar Meher, Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

References

1. Meher, P.K., Sahu, T.K., Rao, A.R. and Wahi, S.D. (2014). A statistical approach for 5' splice site prediction using short sequence motifs and without encoding sequence data. *BMC Bioinformatics*, 15(1), 362.
2. Meher, P.K., Sahu, T.K., Rao, A.R. and Wahi, S.D. (2016). Identification of donor splice sites using support vector machine: a computational approach based on positional, compositional and dependency features. *Algorithms for Molecular Biology*, 11(1), 16.

See Also

`MM1.Feature`, `WAM.Feature`

Examples

```
data(droso)
positive <- droso$positive
negative <- droso$negative
test <- droso$test
pos <- positive[1:200]
neg <- negative[1:200]
tst <- test
enc <- SAE.Feature(positive_class=pos, negative_class=neg, test_seq=tst)
enc
```

`Sparse.Feature` *Nucleotide sequence encoding with 0 and 1.*

Description

In this encoding approach A, T, G and C are encoded as (1,1,1), (1,0,0), (0,1,0) and (0,0,1). This was introduced by Golam Bari *et al.* (2014). Besides, each nucleotide can also be encoded with four bits i.e., A as (1,0,0,0), T as (0,1,0,0), G as (0,0,1,0) and C as (0,0,0,1) as followed in Meher *et al.* (2016).

Usage

```
Sparse.Feature(test_seq)
```

Arguments

`test_seq`	Sequence dataset to be encoded into numeric vector containing 0 and 1, must be an object of class `DNAStringSet`.

Details

Each sequence is encoded independently, without the need of positive and negative classes datasets.

Value

A vector of length $4 * n$ for sequence of n nucleotides long in `test_seq`.

Note

For larger sequence length, high dimensional feature vector will be generated.

Author(s)

Prabina Kumar Meher, Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

References

1. Bari, A.T.M.G., Reaz, M.R. and Jeong, B.S. (2014). Effective DNA encoding for splice site prediction using SVM. *MATCH Commun. Math. Comput. Chem.*, 71: 241-258.
2. Meher, P.K., Sahu, T.K., Rao, A.R. and Wahi, S.D. (2016). A computational approach for prediction of donor splice sites with improved accuracy. *Journal of Theoretical Biology*, 404: 285-294.

Examples

```
data(droso)
test <- droso$test
tst <- test
enc <- Sparse.Feature(test_seq=tst)
enc
```

`Trint.Dist.Feature` *Tri-nucleotide distribution-based encoding of nucleotide sequences.*

Description

This encoding scheme was first time adopted by Wei *et al*. (2013) for prediction of splice sites along with MM1 features. In this encoding technique, distribution of trinucleotides are taken into consideration independently for the exon and intron regions of splice site motifs.

Usage

```
Trint.Dist.Feature(test_seq)
```

Arguments

`test_seq`	Sequence dataset to be transformed into numeric feature vectors. There should be atleat two sequences, must be an object of class `DNAStringSet`.

Details

This encoding scheme is independent of positive and negative datasets. In other words, each sequence can be encoded independently. Further, nucleotide sequence of any length will be transformed into a numeric vector of 64 observations corresponding to 64 combinations of trinucleotides.

Value

A numeric matrix of order $m * 64$, where m is the number of sequences in `test_seq`.

Author(s)

Prabina Kumar Meher, Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

References

Wei, D., Zhang, H., Wei, Y. and Jiang, Q. (2013). A novel splice site prediction method using support vector machine. *J Comput Inform Syst.*, 920: 8053-8060.

Examples

```
data(droso)
test <- droso$test
tst <- test
enc <- Trint.Dist.Feature(test_seq=tst)
enc
```

WAM.Feature *Nucleic acid sequence encoding based on weighted array model.*

Description

Unlike weighted matrix method (WMM), first order nucleotide dependencies are accounted in weighted array model (WAM). The WAM was introduced by Zhang and Marr (1993) for locating splicing signal on nuclotide sequences. The WAM was employed by Meher *et al.* (2016) for encoding of splice site motifs.

Usage

```
WAM.Feature(positive_class, negative_class, test_seq)
```

Arguments

`positive_class`	Sequence dataset of the positive class, must be an object of class `DNAStringSet`.
`negative_class`	Sequence dataset of the negative class, must be an object of class `DNAStringSet`.
`test_seq`	Sequences to be encoded into numeric vectors, must be an object of class `DNAStringSet`.

Details

In this encoding approach, a vector of two observations will be obtained for each sequence, corresponds to the situation when only positive class and both positive & neagtive datasets are used for encoding. This encoding scheme is also invariant to the length of the sequence.

Value

A numeric matrix of order $m * 2$, where m is the number of sequences in `test_seq`.

Author(s)

Prabina Kumar Meher, Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

References

1. Zhang, M. and Marr, T. (1993). A weight array method for splicing signal analysis. *Comput Appl Biosci.*, 9(5): 499-509.
2. Meher, P.K., Sahu, T.K., Rao, A.R. and Wahi, S.D. (2016). Identification of donor splice sites using support vector machine: a computational approach based on positional, compositional and dependency features. *Algorithms for Molecular Biology*, 11(1): 16.

See Also

`MM1.Feature`, `PN.Fdtf.Feature`

Examples

```
data(droso)
positive <- droso$positive
negative <- droso$negative
test <- droso$test
pos <- positive[1:200]
neg <- negative[1:200]
tst <- test
enc <- WAM.Feature(positive_class=pos, negative_class=neg, test_seq=tst)
enc
```

`WMM.Feature` *Weighted matrix model based mapping of nucleotide sequences into vectors of numeric observations.*

Description

The weighted matrix model (WMM) was developed by Staden (1984) for prediction of splice sites. In this technique, position weight matrix are computed from the aligned positive dataset which is then used for computing the liklihood of any nucleotide sequence being a positive or negative class. The position weight matrix was later on used by Meher et al. (2016) for encoding of splice site motifs for prediction using supervised learning models.

Usage

```
WMM.Feature(positive_class, negative_class, test_seq)
```

Arguments

`positive_class` Sequence dataset of the positive class, must be an object of class `DNAStringSet`.

`negative_class` Sequence dataset of the negative class, must be an object of class `DNAStringSet`.

`test_seq` Sequences to be encoded into numeric vector, must be an object of class `DNAStringSet`.

Details

In this encoding approach, a vector of two observations will be obtained for each sequence, corresponds to the situation when only positive class and both positive and neagtive datasets are used for encoding. This encoding scheme is also invariant to the length of the sequence.

Value

A numeric matrix of order $m * 2$, where m is the number of sequences in `test_seq`.

Author(s)

Prabina Kumar Meher, Indian Agricultural Statistics Research Institute, New Delhi-110012, INDIA

References

1. Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research*, 12: 505-519.
2. Meher, P.K., Sahu, T.K., Rao, A.R. and Wahi, S.D. (2016). Identification of donor splice sites using support vector machine: a computational approach based on positional, compositional and dependency features. *Algorithms for Molecular Biology*, 11(1), 16.

See Also

`Bayes.Feature`, `MN.Fdtf.Feature`

Examples

```
data(droso)
positive <- droso$positive
negative <- droso$negative
test <- droso$test
pos <- positive[1:200]
neg <- negative[1:200]
tst <- test
enc <- WMM.Feature(positive_class=pos, negative_class=neg, test_seq=tst)
enc
```

Index