TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

Math Feature Encodings

Sourajyoti Datta
Project – Collaborative Intelligence (DFKI)
Department of Computer Science
Technische Universität Kaiserslautern, Germany

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Major encoding techniques

- Numerical Mapping
- Numerical Mapping and Fourier Transform
- Chaos Game Representation
- Entropy
- Graphs
- K-mer
- Accumulated Nucleotide Frequency (ANF)
- Open Reading Frame (ORF)
- Fickett score
- Pseudo K-tuple nucleotide composition
- Xmer k-Spaced Ymer Composition Frequency (kGap)

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Numerical Mapping

- **Binary**
  - Encodes like one-hot encoding, except the dimension is flattened out as (A + C + T + G).
    - Dimensions: ***L*4*** (L is the length of longest sequence). Zero Padding done at the end, if required.
    - E.g., AACTGT = [1,1,0,0,0,0, 0,0,1,0,0,0, 0,0,0,1,0,1, 0,0,0,0,1,0]
- **Real**
  - Encodes each nucleotide to a fixed real value, keeping sequence intact.
    - Mapping: (**A :** -1.5**, C :** 0.5**, G :** -0.5**, T :** 1.5)
    - Dimensions: **L** (L is the length of longest sequence). Zero Padding done at the end, if required.
    - E.g., AACTGT = [-1.5, -1.5, 0.5, 1.5, -0.5, 1.5]
- **Integer**
  - Encodes each nucleotide to a fixed integer value, keeping sequence intact.
    - Mapping: (**A :** 2**, C :** 1**, G :** 3**, T :** 0)
    - Dimensions: **L** (L is the length of longest sequence). Zero Padding done at the end, if required.
    - E.g., AACTGT = [2, 2, 1, 0, 3, 0]

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Numerical Mapping

- **Electron-Ion Interaction Pseudopotential (EIIP)**
  - Encodes each nucleotide to a fixed real value, keeping sequence intact.
    - Mapping: (**A :** 0.1260**, C :** 0.1340**, G :** 0.0806**, T :** 0.1335)
    - Dimensions: **L** (L is the length of longest sequence). Zero Padding done at the end, if required.
    - E.g., AACTGT = [0.1260, 0.1260, 0.1340, 0.1335, 0.0806, 0.1335]
- **Atomic Number**
  - Encodes each nucleotide to a fixed integer value, keeping sequence intact.
    - Mapping: (**A :** 70**, C :** 58**, G :** 78**, T :** 66)
    - Dimensions: ***L*** (L is the length of longest sequence). Zero Padding done at the end, if required.
    - E.g., AACTGT = [70, 70, 58, 66, 78, 66]
- **Complex Number**
  - Encodes each nucleotide to a fixed complex value, keeping sequence intact.
    - Mapping: (**A :** (1+1j)**, C :** (-1+1j)**, G :** (-1-1j)**, T :** (1-1j))
    - Dimensions: ***L*** (L is the length of longest sequence). Zero Padding done at the end, if required.
    - E.g., AACTGT = [(1+1j), (1+1j), (-1+1j), (1-1j), (-1-1j), (1-1j)]

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Numerical Mapping

- **Z-curve**
  - Encodes each mol nucleotide ecule to three dimensions, and then flattened out.
    - Mapping:

$$\hat{\mathbf{X}}_1(i) = \begin{cases} \mathbf{X}(i-1)+1 & \text{if } X(i) = T \vee G \\ \mathbf{X}(i-1)+(-1) & \text{otherwise} \end{cases}$$

$$\hat{\mathbf{X}}_2(i) = \begin{cases} \mathbf{X}(i-1)+1 & \text{if } X(i) = A \vee C \\ \mathbf{X}(i-1)+(-1) & \text{otherwise} \end{cases}$$

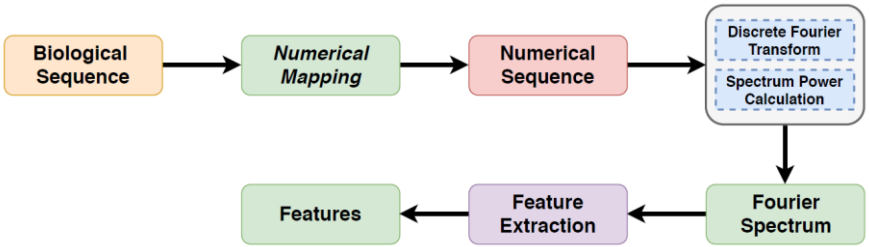
$$\hat{\mathbf{X}}_3(i) = \begin{cases} \mathbf{X}(i-1)+1 & \text{if } X(i) = A \vee T \\ \mathbf{X}(i-1)+(-1) & \text{otherwise} \end{cases}$$

    - Dimensions: ***L*3*** (L is the length of longest sequence). Zero Padding done at the end, if required.
    - E.g., AACTGT = [1,2,1,0,1,0 1,2,3,2,1,0, 1,2,1,2,1,2]

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Numerical Mapping - Fourier Transform

- First, all the numerical mappings are generated as before.
- Then, Discreet Fourier Transform (DFT) is applied using Fast Fourier Transform (FFT)
  - Reveals hidden periodicities after transformation of time domain (sequence) data to frequency domain space.
  - Statistical features are then extracted to generate the dimensions
- Dimensions: ***19***
  - Average
  - Median
  - Maximum
  - Minimum
  - Peak
  - Non-Elevated Peak
  - Sample Standard Deviation
  - Population Standard Deviation
  - 15th Percentile
  - 25th Percentile
  - 50th Percentile
  - 75th Percentile
  - Amplitude
  - Variance
  - Interquartile range
  - Semi interquartile range
  - Coefficient of variation
  - Skewness
  - Kurtosis



TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Chaos Game Representation

- Makes use of Chaos Game and fractals, providing numerical as well as visual representation
- Encodes nucleotide as well as sequence mapping
  - Begin with a square, where each vertex represents a nucleotide. Define an initial point.
  - CGR position of each nucleotide of the DNA sequence is calculated by moving a pointer to half the distance between the previous point and the corner square of the current nucleotide.

- **Classical CGR**
  - The basic CGR representation as explained above.
  - Dimensions: ***L*2*** *(L is the length of longest sequence)*

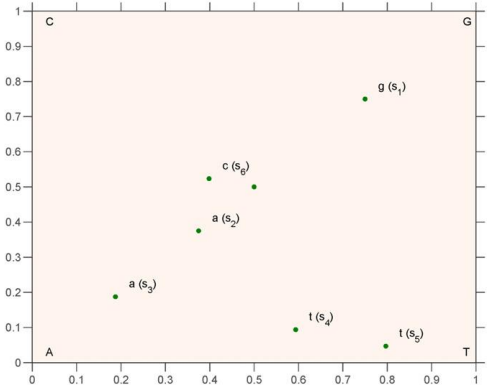


Fig: CGR for "***GAATTC***"

- **Classical CGR - Fourier**
  - Once, the CGR representation is extracted, Discreet Fourier Transform is performed to map from time to frequency domain.
    - Statistical features are then extracted to generate the dimensions
  - Dimensions: ***19***

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Chaos Game Representation

- **Frequency CGR (FCGR)**
  - From the CGR representation, provides a matrix that contains the frequency of the k-mers extracted from the DNA sequences, normalized by the total number of windows
    - Dimensions: **L-k+1** (L is the length of longest sequence, k is length of kmer).

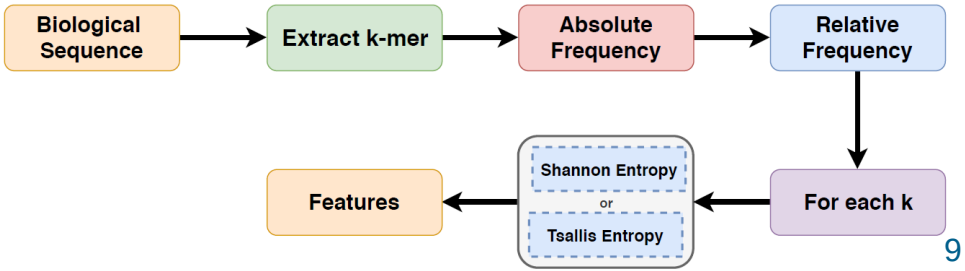
- **Frequency CGR - Fourier**
  - Once, the FCGR representation is extracted, Discreet Fourier Transform is performed to map from time to frequency domain.
    - Statistical features are then extracted to generate the dimensions
  - Dimensions: ***19***

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Entropy

- Entropy is a measure of the uncertainty associated with a probabilistic experiment
- In this method, each sequence is mapped in the frequency of neighboring bases k, generating statistical information.
  - Histograms with short bins are adopted, that occur for k = 1 [{A}, {C}], up to histograms with long sequence counting bins such as k = 7 [{AACCGTG, AGAGAAC}]
  - Generate relative frequencies from absolute frequencies of each ***k***
  - Then the respective entropies are applied
  - Dimensions: ***k*** (k is the maximum length of k-mers)

- **Shannon**
  - Quantifies the amount of information in a variable
  - Reaches a single value that quantifies the information contained in different observation periods (k-mers)
- **Tsallis**
  - Generalized form of the Shannon's entropy



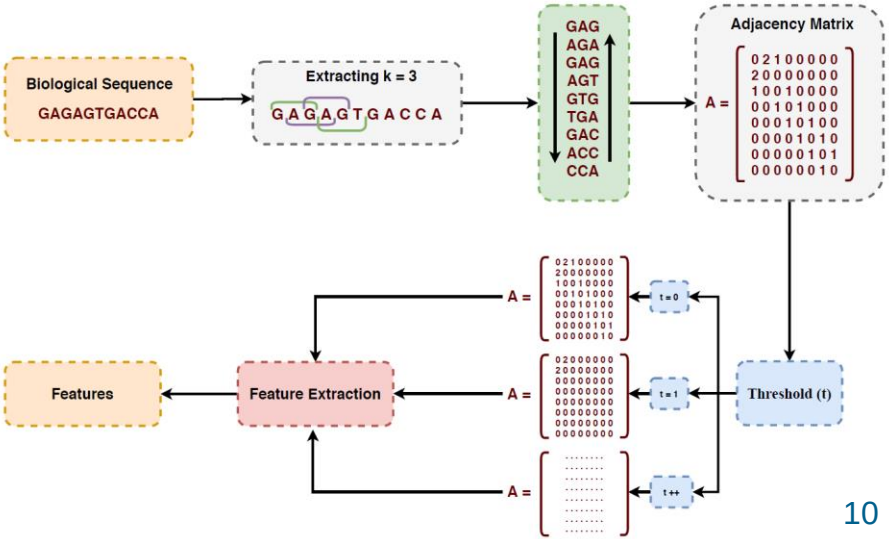
TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Graphs

- Feature extraction model based on complex networks
- Each sequence is mapped in the frequency of neighboring bases k (k = 3)
- This mapping is converted to an undirected graph represented by an adjacency matrix
- Feature extraction is performed using a threshold scheme

- **Complex Networks (with threshold)**
  - Dimension: ***12 * t*** (t is the threshold i.e., the number of subgraphs)

- **Complex Networks (without threshold - v2)**
  - Dimension: ***27 * k*** (k in k-mers)



TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

K-mer

- **Basic K-mer**
  - Generate all k-mers (for k = 1 .. k), for all sequences
  - Calculate their probabilities (i.e., count_of_kmers/windows_in_sequence)
    - Dimensions: ***4^k*** (k is the maximum length of k-mers)

- **Reverse Complement K-mer**
  - Generate all (complement of) k-mers (for k = 1 .. k), for all sequences
    - Complements are:
      - A <-> T
      - C <-> G
      - G <-> C
      - T <-> A
  - Calculate their probabilities (i.e., count_of_kmers/windows_in_sequence)
    - Dimensions: ***4^k*** (k is the maximum length of k-mers)

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

K-mer

- **Nucleic Acid Composition (NAC)**
  - Generate all 1-mers, for all sequences
  - Calculate their probabilities (i.e., count_of_kmers/windows_in_sequence)
    - Dimensions: ***4***

- **Di-nucleotide Composition (DNC)**
  - Generate all 2-mers (for k = 2), for all sequences
  - Calculate their probabilities (i.e., count_of_kmers/windows_in_sequence)
    - Dimensions: ***16***

- **Tri-nucleotide Composition (TNC)**
  - Generate all 3-mers (for k = 3), for all sequences
  - Calculate their probabilities (i.e., count_of_kmers/windows_in_sequence)
    - Dimensions: ***64***

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Accumulated Nucleotide Frequency

- **Accumulated Nucleotide Frequency (ANF)**
  - For each nucleotide in the sequence, their cumulative frequency (nucleotide-wise) is calculated
  - The frequency is standardized by dividing by their individual position in the sequence
    - Dimensions: ***L*** (L is the length of longest sequence, k is length of kmer)
    - Example:
      - Seq: “AAGTAC”
      - Mapping: [1/1, 2/2, 1/3, 1/4, 3/5, 1/6] = [1, 1, 0.333, 0.25, 0.6, 0.167]

- **Accumulated Nucleotide Frequency with Fourier (ANFF)**
  - First, the ANF is generated.
  - Then, Discreet Fourier Transform (DFT) is applied using Fast Fourier Transform (FFT)
    - Reveals hidden periodicities after transformation of time domain (sequence) data to frequency domain space.
    - Statistical features are then extracted to generate the dimensions
    - Dimensions: ***19*** (Similar to the numerical-mapping Fourier transformed dimensions)

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Open Reading Frame (ORF)

- Features for discovering coding sequences
- An ORF is a continuous stretch of codons that begins with a start codon (usually AUG/ATG) and ends at a stop codon (usually UAA/TAA, UAG/TAG or UGA/TGA)
- For each sequence, the first codon is extracted for the lengths (k = 1,2,3), and moving forward, all non-overlapping 3-mers are extracted, for each of the start codons as follows:

1. ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2. A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3. AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A

- Then, the metrics are generated per sequence. Dimensions: ***10***

- Maximum ORF Length
- Minimum ORF Length
- Mean of ORF Length
- Standard deviation of ORF Lengths
- Coefficient of variation of ORF Lengths
- Maximum GC Measure
- Minimum GC Measure
- Mean of GC Measures
- Standard deviation of GC Measures
- Coefficient of variation of GC Measures

* Where, ORF is each possible start/end combination as shown above

* Where, GC Measure is the percentage of G and C nucleotides in each of the ORFs

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Fickett score

- Measures the coding potential based on compositional bias between codon positions by estimating how asymmetric is the distribution of nucleotides at the three triplet positions in the sequence
- Used to evaluate each base's unequal content frequency and asymmetrical distribution in the positions of codons in one sequence
- The Fickett score is computed for both the ORF's and the complete sequence
  - Dimensions: ***2***

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Xmer k-Spaced Ymer Composition Frequency (kGap)

- For every sequence, generates features as the frequency of k-gapped mers characterized by the following parameters:
  - Frequency of ***kgap*** 1 = A_A, 2 = A__A, 3 = A___A...
  - ***Before*** 1 = A_A, 2 = AA_A, 3 = AAA_A...
  - ***After*** 1 = A_A, 2 = A_AA, 3 = A_AAA...

- For all sequences, every possible combination of nucleotides is generated with the above parameters, and their counts as value for the dimension.
  - Dimensions: $4^X * 4^Y$ [X = before, Y=after]