1.

1.1 Global protein sequence descriptor

**Global Protein Sequence Descriptor** là một phương pháp được sử dụng để biểu diễn các trình tự protein dưới dạng số, dựa trên các đặc tính lý hóa của các acid amin. Phương pháp này được phát triển nhằm mục đích mã hóa trình tự protein theo các thuộc tính quan trọng, giúp các mô hình machine learning có thể học và dự đoán các tính chất của protein, như khả năng gấp cuộn (folding) hoặc khả năng kháng khuẩn.

Các đặc điểm chính của Global Protein Sequence Descriptor

1. **Dựa trên các thuộc tính lý hóa (physicochemical properties):**
   * Phương pháp này mã hóa protein dựa trên các thuộc tính lý hóa của các amino acid, như:
     + **Độ kỵ nước (hydrophobicity)**
     + **Thể tích van der Waals chuẩn hóa (normalized van der Waals volume)**
     + **Độ phân cực (polarity)**
     + **Khả năng phân cực (polarizability)**
     + **Điện tích (charge)**
     + **Cấu trúc thứ cấp (secondary structure)**
     + **Khả năng tiếp xúc với dung môi (solvent accessibility)**
   * Mỗi thuộc tính lý hóa này có thể ảnh hưởng đến cấu trúc và chức năng của protein, do đó, mã hóa chúng giúp phản ánh các đặc điểm sinh học của chuỗi protein.
2. **Phân nhóm các amino acid:**
   * Với mỗi thuộc tính lý hóa, các amino acid được chia thành ba nhóm (classes) dựa trên giá trị của thuộc tính đó.
   * Ví dụ, theo thuộc tính **độ kỵ nước (hydrophobicity)**, các amino acid có thể được chia thành:
     + **Polar** (R, K, E, D, Q, N)
     + **Neutral** (G, A, S, T, P, H, Y)
     + **Hydrophobic** (C, L, V, I, M, F, W)
3. **Mã hóa phân bố vị trí (Distribution descriptors):**
   * Phương pháp này tính toán các chỉ số phân bố của các amino acid thuộc từng nhóm, dựa trên tỷ lệ vị trí của các phần tử đầu tiên, 25%, 50%, 75%, và 100% trong chuỗi protein.
   * Chỉ số phân bố này giúp xác định mẫu phân bố của mỗi nhóm amino acid trong chuỗi. Ví dụ, nếu các amino acid kỵ nước phân bố ở đầu và cuối chuỗi, mô hình machine learning có thể nhận diện và sử dụng thông tin này trong dự đoán.
4. **Phương pháp tính toán descriptor:**
   * Mỗi descriptor được tính bằng tỷ lệ vị trí của một amino acid thuộc nhóm nào đó so với tổng chiều dài của chuỗi.
   * Công thức chung: descriptor value=(PL)×100%\text{descriptor value} = \left( \frac{P}{L} \right) \times 100\%descriptor value=(LP​)×100%, trong đó:
     + **P** là vị trí của amino acid trong chuỗi,
     + **L** là tổng số amino acid trong chuỗi.

Ứng dụng của Global Protein Sequence Descriptor

* **Dự đoán cấu trúc và chức năng protein:** Phương pháp này giúp mã hóa thông tin về trình tự protein, hỗ trợ các mô hình machine learning dự đoán các tính chất phức tạp của protein.
* **Xác định peptide kháng khuẩn (AMPs):** Bằng cách mã hóa các đặc tính của AMPs, ta có thể sử dụng machine learning để phân loại và xác định các peptide có khả năng kháng khuẩn.
* **Phân loại protein:** Global Protein Sequence Descriptor có thể giúp phân loại các protein thành các nhóm dựa trên mẫu phân bố đặc trưng của các thuộc tính lý hóa, giúp hiểu rõ hơn về cơ chế sinh học của chúng.

Tóm lại

Global Protein Sequence Descriptor là một phương pháp mã hóa quan trọng, giúp chuyển đổi các trình tự protein thành các giá trị số phản ánh các đặc tính lý hóa của chúng, từ đó giúp cho các mô hình học máy có thể hiểu và xử lý dữ liệu sinh học phức tạp hơn.

**Distribution Descriptor** là một loại chỉ số mô tả trong Global Protein Sequence Descriptor dùng để thể hiện phân bố của các nhóm acid amin theo các thuộc tính lý hóa trong một chuỗi protein. Mỗi nhóm amino acid được xác định dựa trên các tính chất lý hóa cụ thể và có thể chia thành các lớp như **Polar**, **Neutral**, và **Hydrophobic** theo độ kỵ nước. **Distribution descriptor** được tính toán dựa trên vị trí của các phần tử thuộc các lớp này trong chuỗi protein.

### **Phương pháp tính toán Distribution Descriptor**

1. **Lớp tính chất lý hóa:** Đầu tiên, các amino acid trong chuỗi protein được phân loại thành các nhóm theo một tính chất cụ thể. Ví dụ, tính chất **độ kỵ nước** (hydrophobicity) có thể chia amino acid thành các nhóm sau:
   * **Polar (P)**: R, K, E, D, Q, N
   * **Neutral (N)**: G, A, S, T, P, H, Y
   * **Hydrophobic (H)**: C, L, V, I, M, F, W
2. **Xác định phân bố của các phần tử:** Sau đó, chúng ta sẽ tính toán các giá trị phân bố của các phần tử thuộc từng lớp dựa trên các điểm phần trăm:
   * **First residue**: Vị trí xuất hiện đầu tiên của một amino acid thuộc lớp đó.
   * **25% residue**: Vị trí của phần tử thứ 25% trong chuỗi của nhóm amino acid đó.
   * **50% residue**: Vị trí của phần tử thứ 50% trong chuỗi của nhóm amino acid đó.
   * **75% residue**: Vị trí của phần tử thứ 75%.
   * **100% residue**: Vị trí của phần tử cuối cùng của nhóm trong chuỗi.
3. **Công thức tính Distribution Descriptor**:
   * Nếu PPP là vị trí của phần tử được xác định (first, 25%, 50%, 75%, hoặc 100%) và LLL là độ dài chuỗi protein, giá trị distribution descriptor được tính bằng: descriptor value=(PL)×100%\text{descriptor value} = \left( \frac{P}{L} \right) \times 100\%descriptor value=(LP​)×100%

### **Ví dụ tính toán cụ thể**

Giả sử chúng ta có chuỗi peptide sau: **"GLFDIIKKIAESI"** với độ dài 13.

#### **Bước 1: Phân loại amino acid theo độ kỵ nước (hydrophobicity)**

* **Polar (P)**: D, K, E
* **Neutral (N)**: G, A, S
* **Hydrophobic (H)**: L, F, I, I, I

Viết lại chuỗi với ký hiệu cho từng nhóm: **"NHHPHHPPHNPNH"**

#### **Bước 2: Tính toán Distribution Descriptor cho từng nhóm**

**Ví dụ tính Distribution Descriptor cho lớp Polar (P):**

1. **First residue** của lớp Polar (P): Amino acid đầu tiên thuộc lớp Polar là **D**, ở vị trí thứ 4 trong chuỗi.
   * **Descriptor value**: (413)×100%=30.77%\left( \frac{4}{13} \right) \times 100\% = 30.77\%(134​)×100%=30.77%
2. **25% residue** của lớp Polar (P): Vì chuỗi có 4 phần tử thuộc lớp Polar, nên phần tử 25% là phần tử thứ nhất trong lớp (cùng vị trí với phần tử đầu tiên).
   * **Descriptor value**: (413)×100%=30.77%\left( \frac{4}{13} \right) \times 100\% = 30.77\%(134​)×100%=30.77%
3. **50% residue** của lớp Polar (P): Đây là phần tử thứ hai của lớp Polar, xuất hiện ở vị trí **8**.
   * **Descriptor value**: (813)×100%=61.54%\left( \frac{8}{13} \right) \times 100\% = 61.54\%(138​)×100%=61.54%
4. **75% residue** của lớp Polar (P): Đây là phần tử thứ ba của lớp Polar, xuất hiện ở vị trí **9**.
   * **Descriptor value**: (913)×100%=69.23%\left( \frac{9}{13} \right) \times 100\% = 69.23\%(139​)×100%=69.23%
5. **100% residue** của lớp Polar (P): Phần tử Polar cuối cùng xuất hiện ở vị trí **11**.
   * **Descriptor value**: (1113)×100%=84.62%\left( \frac{11}{13} \right) \times 100\% = 84.62\%(1311​)×100%=84.62%

### **Tóm lại**

* Lớp Polar (P) có các distribution descriptors lần lượt là:
  + First residue: 30.77%
  + 25% residue: 30.77%
  + 50% residue: 61.54%
  + 75% residue: 69.23%
  + 100% residue: 84.62%

**Distribution descriptor** giúp biểu diễn thông tin phân bố của các nhóm amino acid trong chuỗi protein. Điều này hữu ích trong các mô hình machine learning, vì nó cung cấp một cách mã hóa định lượng về tính chất của chuỗi protein mà các thuật toán học máy có thể xử lý.

Để mô tả một chuỗi protein thì có 3 descriptor (global)

Composition ( C ): to describe global composition

Transition (T): the frequencies with which the property changes along the entire length of the protein

Distribution (D): the distribution patterns of the property along the sequence

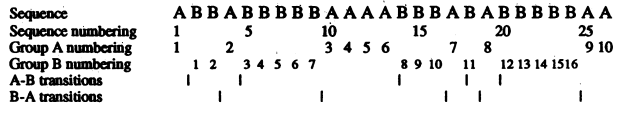
Example:

Bộ mô tả phân bố (Distribution descriptor) giúp biểu diễn thông tin về sự phân bố của các nhóm axit amin trong chuỗi protein. Điều này rất hữu ích trong các mô hình học máy, vì nó cung cấp một cách mã hóa định lượng về tính chất của chuỗi protein mà các thuật toán học máy có thể xử lý.

Để mô tả một chuỗi protein, có 3 bộ mô tả (descriptor) toàn cục:

1. **Composition (C)**: Mô tả thành phần tổng thể của chuỗi protein, tức là tỷ lệ của từng loại axit amin trong chuỗi.
2. **Transition (T)**: Tần suất thay đổi của tính chất (hoặc nhóm axit amin) dọc theo toàn bộ chiều dài của chuỗi protein, cho thấy sự chuyển đổi giữa các trạng thái hoặc nhóm tính chất khác nhau trong chuỗi.
3. **Distribution (D)**: Mô tả các mô hình phân bố của tính chất trong chuỗi protein, tức là cách các axit amin có tính chất đặc biệt phân bố dọc theo chuỗi.

Ví dụ:  
Giả sử chuỗi protein có các đặc tính phân bố khác nhau cho các nhóm axit amin A và B. Bộ mô tả phân bố sẽ xác định các tỷ lệ phần trăm của các dư lượng nhóm A và B trong các đoạn khác nhau của chuỗi protein, chẳng hạn như 25%, 50%, 75% và 100%. Các thông số này sẽ cung cấp thông tin chi tiết về sự phân bố của các đặc tính này trong toàn bộ chuỗi.



Composition descriptor: C

10 type A residues (ni = 10) and 16 type B residues (n2 = 16).

The percent compositions are calculated as follows: n1 x 100.0/(ni + n2) = 38.5% for A and n2 x 100.0/(ni + n2) = 61.5% for B.

Transition descriptor: T

The second descriptor, T, characterizes the percent frequency with which A is followed by B or B is followed by A. In this case, there are 10 transitions of this type, that is (10/25) x 100.0 = 40%.

**Bộ mô tả thành phần: C**  
10 dư lượng loại A (n₁ = 10) và 16 dư lượng loại B (n₂ = 16).  
Tỷ lệ phần trăm thành phần được tính như sau:  
n₁ x 100.0/(n₁ + n₂) = 38.5% cho A và n₂ x 100.0/(n₁ + n₂) = 61.5% cho B.

**Bộ mô tả chuyển tiếp: T**  
Bộ mô tả thứ hai, T, mô tả tần suất phần trăm mà A được theo sau bởi B hoặc B được theo sau bởi A. Trong trường hợp này, có 10 lần chuyển tiếp như vậy, tức là (10/25) x 100.0 = 40%.

Distribution (D):

For a given property of amino acids, the distribution of the property along the protein chain is described by five chain lengths (in percent), within which the first, 25%, 50%, 75%, and 100% of the amino acids with a certain property are contained. In the example of Fig. 2, the first residue of group A coincides with the beginning of the chain, so the first number of D descriptor equals 0.0. Twenty-five percent of all group A residues (rounded to 2 residues) are contained within the first 4 residues of the protein chain, so the second number equals (4/26) x 100.0% = 15.4%. Siniilarly, 50% of group A residues are within the first 12 residues of the chain; thus, the third number is (12/26) x 100.0% = 46.1%. The fourth and fifth numbers of the distribution descriptor are 73.1% and 100%, respectively. Analogous numbers for group B are 7.5%, 23.1%, 53.8%, 79.9%, and 92.3%, respectively. In summary, 13 numbers are used to describe the model sequence shown in Fig. 2 with respect to a given property (type A or B): 2 for composition, 1 for tra-nsition, and 10 for distribution. Thus, the chain descriptor values to be used as the input for NN are 38.5, 61.5, 40.0, 0.0, 15.4, 46.1, 73.1, 100.0, 7.6, 23.1, 53.8, 76.9, and 92.3.

Combined protein sequence descriptors

For each of the chosen amino acids attributes, all three descriptors (C, T, D) were calculated, combined, and used as input parameters for NN training.

Đối với một đặc tính cho trước của các axit amin, sự phân bố của đặc tính đó dọc theo chuỗi protein được mô tả bằng năm độ dài chuỗi (theo phần trăm), trong đó có 25%, 50%, 75%, và 100% các axit amin có đặc tính nhất định nằm trong các độ dài này. Trong ví dụ của Hình 2, dư lượng đầu tiên của nhóm A trùng với đầu chuỗi, vì vậy số đầu tiên của bộ mô tả D bằng 0.0. Hai mươi lăm phần trăm tổng số dư lượng nhóm A (làm tròn thành 2 dư lượng) nằm trong bốn dư lượng đầu tiên của chuỗi protein, vì vậy số thứ hai bằng (4/26) x 100.0% = 15.4%. Tương tự, 50% các dư lượng nhóm A nằm trong 12 dư lượng đầu tiên của chuỗi; do đó, số thứ ba là (12/26) x 100.0% = 46.1%. Số thứ tư và thứ năm của bộ mô tả phân bố lần lượt là 73.1% và 100%. Các số tương tự cho nhóm B là 7.5%, 23.1%, 53.8%, 79.9%, và 92.3%. Tóm lại, 13 số được sử dụng để mô tả chuỗi mẫu trong Hình 2 theo một đặc tính cho trước (loại A hoặc B): 2 cho thành phần, 1 cho chuyển đổi, và 10 cho phân bố. Do đó, các giá trị bộ mô tả chuỗi được sử dụng làm đầu vào cho NN là 38.5, 61.5, 40.0, 0.0, 15.4, 46.1, 73.1, 100.0, 7.6, 23.1, 53.8, 76.9 và 92.3.

Các bộ mô tả chuỗi protein kết hợp  
Đối với mỗi thuộc tính axit amin được chọn, tất cả ba bộ mô tả (C, T, D) đã được tính toán, kết hợp và sử dụng làm tham số đầu vào cho việc huấn luyện NN.

Table: Amino acid attributes and the division of the amino acids into three groups for each attribute

Bảng: Các thuộc tính của amino acid và sự phân chia các amino acid thành ba nhóm cho mỗi thuộc tính

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Divisions** |  |  |
| Hydrophobicity | Polar R, K, E, D, Q, N | Neutral G, A, S, T, P, H, Y | Hydrophobic C, L, V, I, M, F, W |
| Normalized van der Waals volume | Volume range 0-2.78 G, A, S, T, P, D | Volume range 2.95-4.0 N, V, E, Q, I, L | Volume range 4.03-8.08 M, H, K, F, R, Y, W |
| Polarity | Polarity value 4.9-6.2 L, I, F, W, C, M, V, Y | Polarity value 8.0-9.2 P, A, T, G, S | Polarity value 10.4-13 H, Q, R, K, N, E, D |
| Polarizability | Polarizability value 0-0.108 G, A, S, D, T | Polarizability value 0.128-0.186 C, P, N, V, E, Q, I, L | Polarizability value 0.219-0.409 K, M, H, F, R, Y, W |
| Charge | Positive K, R | Neutral A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V | Negative D, E |
| Secondary structure | Helix E, A, L, M, Q, K, R, H | Strand V, I, Y, C, W, F, T | Coil G, N, P, S, D |
| Solvent accessibility | Buried A, L, F, C, G, I, V, W | Exposed P, K, Q, E, N, D | Intermediate M, P, S, T, H, Y |

The division is based on the clusters of the amino acid indices of Tomii and Kanehisa ( 5 , 40 ) for each of the seven attributes. For such attributes as secondary structure and solvent accessibility, the division is based on statistical appearance of each amino acid in a specific state.

Sự phân chia này dựa trên các nhóm chỉ số amino acid của Tomii và Kanehisa (5, 40) cho mỗi thuộc tính trong bảy thuộc tính khác nhau. Đối với những thuộc tính như cấu trúc bậc hai và khả năng tiếp xúc với dung môi, sự phân chia được thực hiện dựa trên sự xuất hiện thống kê của mỗi amino acid trong một trạng thái cụ thể.

