2007－BIO－10 (11)
2007／9／14

# アライメントアルゴリズムの改良

中村 卓　　佐藤 圭子
東京理科大学理工学部情報科学科

**概要** 動的計画法を用いたアライメントは，配列間の差異を最小にする（あるいは配列間のスコアを最大にする）ようなアライメント結果を得ることができる．しかし，そのアライメント結果が複数生じる場合があり，その結果から最も確からしいと思われる一つを選択する方法はまだ確定していない．また，各アミノ酸間（あるいは塩基間）やアミノ酸とギャップ間の差異の定義の仕方によって，得られる結果はもちろん異なってくる．したがって，その差異をどのように決定し，かつ適切なアライメント結果を導き出すかが重要となる．そこで，ラマチャンドラン・プロットを基にして，各アミノ酸間の差異を 3 通りの方法で定義し，それらと，BLOSUM 行列とを組み合わせることで，タンパク質の立体構造と，各アミノ酸のペアの起こりやすさを考慮した差異行列を作成した．

# Improvement of alignment algorithm

Takumi Nakamura and Keiko Sato
Tokyo University of Science, Department of Information Sciences

**Abstract** The sequence alignment based on dynamic programming has been done on a principle to give the shortest difference between two sequences (or the maximum score between two sequences). We have a lot of the alignment results having the same value to the difference; however we have not known a criterion which result is most proper. The alignment result is strongly influenced by the way to define the difference between two amino acids (resp., nucleotides) and between an amino acid (resp., a nucleotide) and a gap. Therefore it is important that we decide how to define the difference and find the proper result for the alignment. In this paper, we define the difference between amino acids in three ways on the basis of Ramachandran plot, then we combine these methods with the BLOSUM matrix. That is, we make the difference matrix taking account of the peptide structure and the frequencies of substitution of amino acids in each position.

## 1. 序章

現在，アライメントを行う際に最も広く使われているスコア行列が BLOSUM 行列である．しかしながら，BLOSUM 行列はアミノ酸置換行列であるため，タンパク質の立体構造

を考慮に入れてはいない．タンパク質の働きは立体構造によって決まるので，立体構造を考慮することは重要である．そこで，ラマチャンドラン・プロットを応用して，タンパク質の立体構造を考慮した行列を作り，それらの行列と BLOSUM 行列を組み合わせることで立体構造と置換の起こりやすさを考慮した行列を作成した．

本論文では，まずラマチャンドラン・プロットから各アミノ酸間の差異を 3 通りの方法で定義した後，それらと BLOSUM 行列とを組み合わせる方法を述べ、作成した差異行列をアライメントアルゴリズムに適用した結果を示す．

## 2. ラマチャンドラン・プロットの応用

### 2.1. ラマチャンドラン・プロット

タンパク質の主鎖の構成する結合部分は様々な角度に折れ曲がっており多くのコンホメーションをとる．しかし，実際に自由に回転できる結合部分は，アミノ基と中心炭素のまわりの回転角 $\phi$ と，カルボキシル基と中心炭素のまわりの回転角 $\psi$ である．$\phi$ と $\psi$ は原子同士が衝突しない無理のない角度に限定されている．実際に $\phi$ と $\psi$ をそれぞれ $x, y$ 軸方向に取りプロットしたものを，この角度を最初に計算した生物学者 G.N.Ramachandran にならって，ラマチャンドラン・プロットと呼ぶ[1]．$\phi$ と $\psi$ の角度分布は側鎖が関係しているので，側鎖の種類，つまりアミノ酸の種類によって異なっている．側鎖が小さいグリシンは，かなり自由な角度を取ることができる（図 1）．しかし，側鎖が比較的大きいプロリンでは，あまり自由な角度を取ることができない（図 2）．

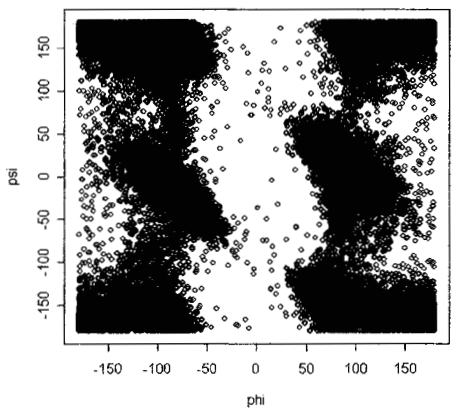


図 1 gly のラマチャンドラン・プロット

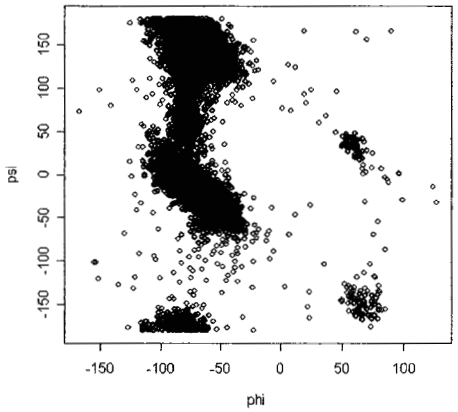


図 2 pro のラマチャンドラン・プロット

### 2.2 アミノ酸間の差異の定義

実際，どのようにラマチャンドラン・プロットからアミノ酸間の差異を計算するか，その方法を説明する．今回使用したアミノ酸の $(\phi, \psi)$ 角度のデータはDASSD (http://www.cs.rmit.edu.au/dassd/) から収集した[2].

まず，ラマチャンドラン・プロットの$\phi$方向，$\psi$方向をそれぞれ等間隔に区切る．これによって作られる升目の中に，いくつデータが入っているかをカウントする．よって，$\phi$方向，$\psi$方向にそれぞれ$N$に分割したとすると，$N\times N$個の数が得られることになる．今回は$N=72$とした[3].

[方法 1]

この方法では，2つのアミノ酸の角度分布の差の絶対値をもって差異とすることを考える．アミノ酸の種類によってプロットの総数が異なる場合も考慮しつつ，分布間の差の絶対値を計算し，さらにその結果が 0 から 1 の範囲に収まるように正規化を行ったものを，各アミノ酸間の差異とする．アミノ酸$a,b$の差異を$d'_{Rama_1}(a,b)$と表すと，次の式で定義できる．

$$d'_{Rama\_1}(a,b)=\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{n_{ij}^{a}}{N_a}-\frac{n_{ij}^{b}}{N_b}\right|}{2}$$

ただし，

$$N_a=\sum_{i=1}^{N}\sum_{j=1}^{N}n_{ij}^{a}\ ,\ N_b=\sum_{i=1}^{N}\sum_{j=1}^{N}n_{ij}^{b}$$

ここで，$n_{ij}^{a},n_{ij}^{b}$は，それぞれアミノ酸$a,b$のラマチャンドラン・プロットにおいて，$\phi$方向の$i$番目，$\psi$方向の$j$番目の升目にプロットされている点の個数である．

[方法 2]

この方法では，$n_{ij}^{a},n_{ij}^{b}$を，それぞれ$N\times N$次元ベクトルの成分とみなす．そして，それぞれを単位ベクトル化した後で，ベクトルの差の絶対値をもって角度分布の差異とすることを考える．その結果が 0 から 1 の範囲に収まるように正規化を行ったものを各アミノ酸間の差異とする．この方法によるアミノ酸$a,b$の差異を$d'_{Rama_2}(a,b)$と表すと，次の式で定義できる．

$$d'_{Rama\_2}(a,b)=\frac{\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{n_{ij}^{a}}{N'_a}-\frac{n_{ij}^{b}}{N'_b}\right)^{2}}}{\sqrt{2}}$$

ただし，

$$N'_a=\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(n_{ij}^{a}\right)^{2}}\ ,\ N'_b=\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(n_{ij}^{b}\right)^{2}}$$

[方法 3]

この方法では，相対エントロピーを使用して，角度分布間の差異とすることを考える．$p=\left\{p_{ij}^{a}\right\}_{i,j=1}^{N}, q=\left\{q_{ij}^{b}\right\}_{i,j=1}^{N}$ $\left(p_{ij}^{a}\equiv\frac{n_{ij}^{a}}{N_a}, q_{ij}^{b}\equiv\frac{n_{ij}^{b}}{N_b}\right)$ をそれぞれアミノ酸 $a,b$ のラマチャンドラン・プロットにおいて，$\phi$ 方向の $i$ 番目，$\phi$ 方向の $j$ 番目の升目にプロットされている点の個数の確率分布とする．その結果が 0 から 1 の範囲に収まるように正規化を行ったものを各アミノ酸間の差異とする．この方法によるアミノ酸 $a,b$ の差異を $d'_{Rama\_3}(a,b)$ と表し，次の式で定める．

$$d'_{Rama\_3}(a,b)=S(p|r)+S(q|r)$$

ただし，

$$r_{ij}=\frac{p_{ij}^{a}+q_{ij}^{b}}{2}$$

とし，

$$S(p|r)=\sum_{i=1}^{N}\sum_{j=1}^{N}p_{ij}^{a}\log\frac{p_{ij}^{a}}{r_{ij}}, S(q|r)=\sum_{i=1}^{N}\sum_{j=1}^{N}q_{ij}^{b}\log\frac{q_{ij}^{b}}{r_{ij}}$$

なお，方法 1 で作成した差異行列を，Ramachandran_1 行列，方法 2 で作成した差異行列を，Ramachandran_2 行列，方法 3 で作成した差異行列を，Ramachandran_3 行列とする．

## 3. BLOSUM行列の応用

### 3.1. BLOSUM行列

BLOSUM 行列とは，類縁配列をマルチプルアライメントし，その結果の中でも特に保存率の高い部分での，アミノ酸の置換の起こりやすさを数値化した行列である．この時の類縁配列の違いから，複数の BLOSUM 行列が存在している．BLOSUM 行列のスコアはプラス値ほど，そのアミノ酸の組み合わせが起こりやすいことを示している．本研究では類似度が 62%以上の配列から数値化した BLOSUM62 行列を使用した．BLOSUM62 行列のスコアの最大は 11 であり，トリプトファンとトリプトファンの組み合わせに与えられている．逆に，一番起こりにくい組み合わせとして，トリプトファンとプロリン，ロイシンとグリシン，等いくつかの組み合わせがあり，スコアは最小の－4 となる[4].

### 3.2. BLOSUM行列の変形

BLOSUM 行列を，前章で作成した 3 つの差異行列と組み合わせるため，BLOSUM 行列を変形したものを transformationalBLOSUM 行列と呼ぶことにする．その

transformationalBLOSUM 行列の各成分となるアミノ酸 $a,b$ の差異 $d'_{transBLO}(a,b)$ は次のように表すことができる.

$$d'_{transBLO}(a,b)=1-\frac{B(a,b)+4}{15}$$

$B(a,b)$ は，アミノ酸 $a,b$ によって決まる BLOSUM 行列の値とする．この変形により，transformationalBLOSUM 行列は BLOSUM 行列で一番スコアの高いトリプトファンとトリプトファンの組み合わせが差異 0 となり，スコアが一番低い組み合わせが差異 1 となり，すべての組み合わせが 0 から 1 の範囲に収まる.

### 3.3. BLOSUM行列とラマチャンドラン・プロットの組み合わせ方法

transformationalBLOSUM 行列の値と，前章で導き出したラマチャンドラン・プロットの差異行列を組み合わせて，新たな差異行列を作成するため，アミノ酸 $a,b$ の差異を次の $d'(a,b)$ で定義した.

$$d'(a,b)=d'_{transBLO}(a,b)\times\alpha+d'_{Rama\_1}(a,b)\times(1-\alpha)$$

ここで，$\alpha$ は，0 から 0.1 刻みで増やしていき，1 までをとることにする．したがって，9 パターンの組み合わせた行列が作成できる．これと同じことを，Ramachandran_2 行列，Ramachandran_3 行列の値を用いて行うため，計 27 パターンの行列が出来上がる.

なお，今後 $\alpha=0.1$ とした時の行列を 0.1＊BLO+0.9＊Rama_1 と表すことにする.

## 4. アライメントアルゴリズム

本研究で用いたアライメントアルゴリズムは，動的計画法を用いて作成された MOU アライメントを改良したペアワイズアライメントアルゴリズムを使用した[5]．MOU アライメントは，配列間の差異を最小とするグローバルアライメントアルゴリズムなので，$D[i,j]$ を配列 $\alpha$ の $i$ 番目と配列 $\beta$ の $j$ 番目の位置での配列間の差異を表すとすると，本研究で使用したアライメントアルゴリズムを次の式で表すことができる.

$$D[i,j]=\min\begin{cases}D[i,j-1]+w\\ D[i-1,j-1]+d(a,b)\\ D[i-1,j]+w\end{cases}$$

ただし，

$$D[i,0]=w\times i\quad,\quad D[0,j]=w\times j$$

とする．$w$ はギャップに対する重みで，$d(a,b)$ はアミノ酸 $a,b$ の差異で，2 章で定義した 3 つの差異行列と，transformationalBLOSUM 行列と，3.3 節で述べた 27 個の差異行列の値をとる．そのため，アライメントアルゴリズムに用いる差異行列は計 31 パターンとなる.

## 5. 実験と結果

まず，使用する配列をNCBIのWebサイト（http://www.ncbi.nlm.nih.gov/）から10種のタンパク質で異なる生物種のアミノ酸配列を10本ずつ収集した（表1）．

### 5.1. 実験方法

まず，1つのタンパク質につき，10本の配列を総当りでペアワイズアライメントを行う．つまり，アライメントを行う配列ペアは45組である．それを10種すべてのタンパク質で行うため，配列ペアは450組になる．その450組で，31すべての差異行列を用いてアライメントを行った．なお，アミノ酸差異とのバランスをとる意味と，ギャップに対する重みに左右されずに，差異行列の検証を行うため，ギャップに対する重みを，1.0，1.5，2.0，2.5と4つの値で試した．

動的計画法を用いるペアワイズアライメントは最も確実なアライメント法ではあるが，アライメント結果が多数導き出されてしまうことがある．しかし，進化の過程を考えるとアライメント結果は1つであるべきである．そこで，全ての差異行列を用いたアライメント結果の中で，1つに絞れた数が多い差異行列をより確からしい差異行列であるとした．

表1　使用配列表

| タンパク質 | accession number | 配列名 | タンパク質 | accession number | 配列名 |
|---|---|---|---|---|---|
| cutinase | XP 755273 | Aspergillus fumigatus AF293 | myoglobin | BAD23846 | Auxis rochei |
| | BAA07428 | Aspergillus oryzae | | NP 776306 | Bos taurus |
| | CAA93255 | Botryotinia fuckeliana | | AAB54102 | Chionodraco rastrospinosus |
| | AAL38030 | Glomerella cingulata | | AAK49781 | Hemitripterus americanus |
| | AAZ95012 | Monilinia fructicola | | NP 976312 | Homo sapiens |
| | CAA46582 | Mycosphaerella rabiei | | NP 038621 | Mus musculus |
| | AAB05922 | Nectria ipomoeae | | AAC69245 | Notothenia coriiceps |
| | CAA61622 | Phytophthora capsici | | AAS92621 | Ochotona curzoniae |
| | AAX55266 | Phytophthora infestans | | NP 067599 | Rattus norvegicus |
| | CAB40372 | Pyrenopeziza brassicae | | NP 999401 | Sus scrofa |
| cytochrome c | AAB72175 | Arabidopsis thaliana | profilin | AAD29409 | Apium graveolens |
| | P 00011 | Canis familiaris | | AAG10091 | Arabidopsis thaliana |
| | CAB16954 | Chlamydomonas reinhardtii | | CAD10376 | Capsicum annuum |
| | CAA25046 | Gallus gallus | | CAD9266 | Cucumis melo |
| | AAR30955 | Helianthus annuus | | CAD10377 | Lycopersicon esculentum |
| | NP 061820 | Homo sapiens | | CAA57632 | Nicotiana tabacum |
| | CAA25899 | Mus musculus | | CAA54686 | Phleum pratense |
| | CAA29050 | Neurospora crassa | | AAD29411 | Prunus avium |
| | AAS67288 | Pichia pastoris | | AAD29410 | Pyrus communis |
| | BAC54258 | Rosellinia necatrix | | AAB22843 | Strongylocentrotus purpuratus |
| ferritin | AAL55388 | Artemia franciscana | rubredoxin | NP 879193 | Bordetella pertussis Tohama |
| | NP 281795 | Campylibacter jejuni | | YP 107689 | Burkholderia pseudomallei |
| | CAB72315 | Daphnia pulex | | CAA09017 | Clostridium butyricum |
| | YP 050566 | Erwinia carotovora | | CAB41597 | Clostridium cellulolyticum |
| | AAQ54714 | Ixodes scapularis | | CAA09015 | Clostridium diolis |
| | AAZ07716 | Puccinellia tenuiflara | | AAK08075 | Desulfovibrio gigas |
| | AAV76910 | Salmonella enterica | | BAA11175 | Desulfovibrio vulgaris |
| | YP 041358 | Staphylococcus aureus | | YP 447487 | Methanosphaera stadtmanae |
| | AAB20316 | Xenopus laevis | | NP 254037 | Pseudomonas aeruginosa |
| | NP 405350 | Yersinia pestis C092 | | NP 623712 | Thermoanaerobacter tengcongensis |
| glucagon | AAB28788 | Amia calva | thioredoxin | NP 281358 | Campylibacter jejuni |
| | NP 776341 | Bos taurus | | NP 418228 | Escherichia coli |
| | NP 001003044 | Canis familiaris | | AAB93304 | Eubacterium acidaminophilum |
| | AAT00451 | Capra hircus | | NP 003320 | Homo sapiens |
| | CAA68827 | Gallus gallus | | CAB40815 | Listeria monocytogenes |
| | AAP35459 | Homo sapiens | | CAC30771 | Mycobacterium leprae |
| | NP 032126 | Mus musculus | | NP 326538 | Mycoplasma pulmonis |
| | AAB28397 | Petromyzon marinus | | NP 446252 | Rattus norvegicus |
| | NP 036839 | Rattus norvegicus | | NP 625184 | Rhodopirellula baltica |
| | NP 999489 | Sus scrofa | | YP 218808 | Salmonella enterica |
| insulin | AAA37041 | Cavia porcellus | ribonuclese P | YP 093879 | Bacillus licheniformis |
| | CAC20109 | Danio rerio | | NP 418159 | Escherichia coli |
| | INEL | Elephantidae | | YP 248685 | Haemophilus influenzae |
| | AAM76640 | Gorilla gorilla | | NP 965835 | Lactobacillus johnsonii |
| | INHY | Cricetinae | | CAD65752 | Lactobacillus plantarum |
| | AAA59172 | Homo sapiens | | NP 487453 | Nostoc |
| | AAA40590 | Octodon degus | | YP 376016 | Pelodictyon luteolum |
| | AAA19033 | Oryctolagus cuniculus | | NP 734911 | Streptococcus agalactiae |
| | AAB60625 | Ovis aries | | ZP 00993220 | Vibrio splendidus |
| | NP 062002 | Rattus norvegicus | | NP 995270 | Yersinia pestis |

### 5.2. 結果

全てのアライメント結果の中で，結果が1つに絞れた確率を行列ごとに示しておく（表2）．この中で最も確からしい差異行列は，transformationalBLOSUM行列とRamachandran_3行列を，3：7の比率で足し合わせた0.3＊BLO+0.7＊Rama_3行列である．

また，各タンパク質について最も確からしい差異行列と，その差異行列を使用したアライメントの結果が1つに絞れた確率を示し，比較のためにtransformationalBLOSUM行列の1つに絞れた確率も示しておく（表3）．

表2　アライメント結果が1つに絞れた確率

| 差異行列 | | 差異行列 | | 差異行列 | |
|---|---|---|---|---|---|
| Ramachandran_1 | 56.11% | Ramachandran_2 | 56.94% | Ramachandran_3 | 57.17% |
| 0.1*BLO+0.9*Rama_1 | 59.06% | 0.1*BLO+0.9*Rama_2 | 58.83% | 0.1*BLO+0.9*Rama_3 | 58.61% |
| 0.2*BLO+0.8*Rama_1 | 56.22% | 0.2*BLO+0.8*Rama_2 | 55.94% | 0.2*BLO+0.8*Rama_3 | 57.22% |
| 0.3*BLO+0.7*Rama_1 | 55.56% | 0.3*BLO+0.7*Rama_2 | 57.28% | 0.3*BLO+0.7*Rama_3 | 59.11% |
| 0.4*BLO+0.6*Rama_1 | 56.22% | 0.4*BLO+0.6*Rama_2 | 55.78% | 0.4*BLO+0.6*Rama_3 | 55.78% |
| 0.5*BLO+0.5*Rama_1 | 55.78% | 0.5*BLO+0.6*Rama_2 | 55.94% | 0.5*BLO+0.6*Rama_3 | 56.72% |
| 0.6*BLO+0.4*Rama_1 | 55.83% | 0.6*BLO+0.4*Rama_2 | 56.56% | 0.6*BLO+0.4*Rama_3 | 56.00% |
| 0.7*BLO+0.3*Rama_1 | 57.11% | 0.7*BLO+0.3*Rama_2 | 55.89% | 0.7*BLO+0.3*Rama_3 | 57.94% |
| 0.8*BLO+0.2*Rama_1 | 56.67% | 0.8*BLO+0.2*Rama_2 | 55.50% | 0.8*BLO+0.2*Rama_3 | 56.67% |
| 0.9*BLO+0.1*Rama_1 | 57.06% | 0.9*BLO+0.1*Rama_2 | 56.94% | 0.9*BLO+0.1*Rama_3 | 56.00% |
| transformationalBLOSUM | 38.89% | | | | |

表3　各タンパク質で最も確からしい差異行列とtransformationalBLOSUM行列のアライメント結果が1つに絞れた確率

| タンパク質 | 差異行列 | 確率 | 差異行列 | 確率 |
|---|---|---|---|---|
| cutinase | 0.1*BLO+0.9*Rama_1 | 51.67% | transformationalBLOSUM | 11.67% |
| cytochrome c | 0.7*BLO+0.3*Rama_3 | 84.44% | transformationalBLOSUM | 49.44% |
| ferritin | 0.3*BLO+0.7*Rama_2 | 70.56% | transformationalBLOSUM | 24.44% |
| glucagon | 0.1*BLO+0.9*Rama_1 | 42.78% | transformationalBLOSUM | 37.78% |
| insulin | 0.1*BLO+0.9*Rama_3 | 57.78% | transformationalBLOSUM | 29.44% |
| myoglobin | 0.7*BLO+0.3*Rama_1 | 75.56% | transformationalBLOSUM | 57.78% |
| profilin | 0.9*BLO+0.1*Rama_1 | 78.33% | transformationalBLOSUM | 61.67% |
| ribonuclese P | Ramachandran_3 | 76.67% | transformationalBLOSUM | 37.22% |
| rubredoxin | 0.9*BLO+0.1*Rama_3 | 64.44% | transformationalBLOSUM | 57.78% |
| thioredoxin | 0.1*BLO+0.9*Rama_2 | 58.89% | transformationalBLOSUM | 21.67% |

## 6. 考察

本実験の結果から，0.3＊BLO+0.7＊Rama_3行列が最も確からしい差異行列であるということは記した．また，各タンパク質もribonuclese PのRamachandran_3行列を用いたアライメントアルゴリズムが最も確からしいという結果となっているのを除いて，transformationalBLOSUM行列とラマチャンドラン・プロットから作成した差異行列を組み合わせた差異行列を用いたアライメントアルゴリズムが最も確からしい結果となっている．さらに，表2，表3は，ギャップに対する重みごとに詳細を述べてはいないが，ギャッ

プに対する重みごとにアライメント結果を見てみると，全てのアライメント結果の中で最も確からしいアライメントアルゴリズムになったのは，0.1＊BLO+0.9＊Rama_2 行列を使用し，ギャップに対する重みを 2.5 としたアルゴリズムで，アライメント結果が 1 つに絞れた確率は 62.0%にもなる．各タンパク質ごとにアライメント結果を見てみると，cytochrome c に対して 0.9＊BLO+0.1＊Rama_2 行列を使用し，ギャップに対する重みを 1.5 としたアルゴリズムが，1 つに絞れた確率が 91.11%にもなり，最も高い結果となっている．

これらの結果は，本研究で作成した差異行列が BLOSUM 行列より優れているということを表している．transformationalBLOSUM 行列とラマチャンドラン・プロットから作成した差異行列を組み合わせた差異行列が確からしい結果となりやすいのは，アミノ酸の置換の起こりやすさとタンパク質の立体構造を考慮に入れているからであろう．

**参考文献**


[1] G. N. Ramachandran, C. Ramakrishnan, V. Sasisekharan “Stereochemistry of polypeptide chain configurations” *J. Mol. Biol.* Vol.7, pp.95-99 (1963)

[2] Saravan Dayalan, Nalaka Dilshan Gooneratne, Savitri Bevinakoppa, Heiko Schoroder “Dihedral angle and secondary structure database of short amino acid fragments” *Bioinformation,* Vol.1(3), pp.78-80 (2006)

[3] T.Kuroiwa, M.Ohya “On Multiple Alignment of Amino Sequences with Protein Structure” *IEICE technical report,* Vol.98, No.211, pp.55-60 (1998)

[4] Steven Henikoff, Joria G. Henikoff, “Amino acid substitution matrices from protein blocks” *Proc. Natl. Acad. Sci.,* Vol.89, pp.10915-10919 (1992)

[5] M.Ohya, S.Miyazaki, Y.Ohshima “A new method of Alignment of Amino Acid Sequences” *Viva Origino* 17, pp.139-151 (1989)