Trabalho Prático 3
Universidade Federal de Minas Gerais
Bioinformática

Alunos: Gustavo Campos Ferreira Guimarães
Osvaldo Luís Henriques de Morais Fonseca

1. INTRODUÇÃO

Neste trabalho, iremos utilizar informações estruturais sobre a triose fosfato isomerase estudada no TP1, a proteína selvagem (2YPIA) e as outras proteínas da família, para tentar identificar as mutações que impactaram na perda da função da proteína mutada (dTIM).

Vamos utilizar os conceitos de contatos estudados no TP2 para criar grafos que representem cada uma dessas proteínas. Os vértices dos grafo representam os resíduos da proteína e as arestas os contatos. A partir desses grafos, podemos utilizar algumas métricas muito usadas em redes complexas: *betweenness, closeness* e grau. Essa métricas podem oferecer informações muito interessantes sobre a importância de um determinado resíduo, consequentemente no impacto que aquela mutação pode causar. Por exemplo, se o resíduo mutado tem um grau muito elevado, isso significa que ele faz contato com muitos outros resíduos, e provavelmente essa mutação pode ocasionar a perda da função.

No TP1, a partir dos alinhamentos de sequências realizados, selecionamos aquelas mutações que pareciam ser mais impactantes, escolhendo mutações que ocorriam pouquíssimas vezes nas sequências da família. Depois, analisamos as características dos resíduos, como: polarização, carga, tamanho da cadeia lateral, etc. A partir dos experimentos, chegamos nas dez mutações possivelmente mais relevantes. Utilizando as informações de estrutura, podemos confirmar ou descartar a relevância de

algumas delas.

2. SOLUÇÃO PROPOSTA

2.1. Tratamento da Base de Dados

Como vimos no TP2, os arquivos do PDB, com as informações estruturais das proteínas, podem conter alguns problemas. Apesar de seguirem um certo padrão, com colunas específicas para cada um dos atributos, pode acontecer das informações de dois atributos concatenarem. Sem espaço entre os atributos, a tarefa de identificá-los é dificultada. O que fizemos foi utilizar os números das colunas para identificar os atributos. Por exemplo, da coluna 1 até a 6 conseguimos obter o atributo que diz se aquela linha contém informações sobre um átomo.

2.2. Implementação

A primeira tarefa a ser realizada é fazer a leitura do arquivo no formato pdb, e armazenar as informações dos resíduos. Como o estudo será feito no nível de resíduo, para simplificar, escolhemos utilizar a posição do carbono alfa dos resíduos para representá-los. Então, ao ler o arquivo, iremos armazenar apenas as informações referentes aos carbonos alfa, e serão armazenados dados como: id do resíduo, id do átomo, posição tridimensional, nome do resíduo, etc. Todas as outras linhas, que não estão relacionadas aos resíduos das proteínas, são descartadas.

Além de armazenas apenas os carbonos alfa, estamos interessados nos resíduos que pertencem a cadeia A. Portanto, o algoritmo só irá utilizar os dados dos resíduos referentes à essa cadeia.

Depois disso, calculamos a distância de cada um dos resíduos a todos os outros da proteína, utilizando como métrica a distância euclideana. Como já foi mencionado, no cálculo de distância usamos a posição dos carbonos alfa de cada resíduo. Com a matriz

de distâncias calculada, podemos definir os contatos. Estamos utilizando para isso uma abordagem baseada em *cut-off*, que utiliza como métrica a distância euclideana. Para um determinado limiar de distância *y*, temos um contato entre dois resíduos se a distância euclideana entre eles for menor que esse limiar. Logo, o que fazemos é criar uma matriz quadrada com os contatos (matriz_contatos[i][j]) e, ao percorrer a matriz de distâncias, se a distância entre o resíduo *i* com o *j* for menor que o limiar *y*, fazemos matriz_contatos[i][j] = 1, e fazemos matriz_contatos[i][j] = 0 caso contrário.

Desta forma, com a matriz de contatos construída, nós vamos gerar um arquivo no formato .net, que contém as informações necessárias para gerar o grafo utilizando a biblioteca *igraph* do pacote estatístico *R*. No arquivo, primeiro temos que colocar os vértices, que são os resíduos, e seus respectivos *labels*. Para isso basta percorrer a estrutura que contém todos os resíduos e suas informações. O início do arquivo fica da seguinte forma:

*Vertices 3
1 “ALA(2)”
2 “ARG(3)”
3 “THR(4)”

Depois devemos definir as arestas, que representam os contatos. Para isso basta percorrer a matriz de contatos, e se na posição *matriz_contatos[i][j]* for igual a 1, então imprime no arquivo da seguinte forma:

*Edges
1 2 2.2342
1 3 1.2323
2 3 3.1231

Agora basta carregar o grafo no *R* para poder realizar as análises do trabalho, podendo utilizar métricas de redes complexas como: *betweenness, closeness* e grau.

3. EXPERIMENTOS

Foram feitos alguns experimentos com a finalidade de obter mais informações sobre as mutações, e cada um deles será detalhado e discutido em uma subseção.

3.1 Alinhamento

Dispostos os grafos obtidos através do cálculo de contatos, possuímos um arcabouço de dados que podem então ser usados para ajudar na decisão de classificar as mutações entre as proteínas propostas. A fim de realizar essa classificação precisamos, portanto, obter a lista das mutações entre a dTIM e a 2YPI.

Como sabemos, o alinhamento sequencial de proteínas não é uma ciência exata. Pequenos ajustes paramétricos resultam em resultados bastante diferentes, de forma que não se pode garantir totalmente o grau de certeza da resposta obtida. Sabendo disso, foram feitos vários alinhamentos, e foi usado um critério de análise subjetivo para decidir o mais adequado. Em outras palavras, observou-se os resultados dos alinhamentos e escolheu-se o que parecia visualmente mais correto.

Os parâmetros e o resultado do alinhamento obtido podem ser visualizados abaixo:

```
Algoritmo de Alinhamento: Biopython.pairwise2
Penalidade por Gap: -10
Penalidade por continuidade de Gap: -0.5
Matriz de Pontuação: PAM250
```

```
Alinhamento obtido:

2YPY: MARTFFVGGNFKLNGSKQSIKEIVERLNTASIPENVEVVICPPATYLDYSVSLVKKPQVTVGAQNAYLKASGAFTGENSV
dTIM: MARTPFVGGNWKMNGTKAEAKELVEALK-AKLPDDVEVVVAPPAVYLDTAREALKGSKIKVAAQNCYKEAKGAFTGEISP
2YPI: DQIKDVGAKWVILGHSERRSYFHEDDKFIADKTKFALGQGVGVILCIGETLEEKKAGKTLDVVERQLNAVLEEV-KDWTN
dTIM: EMLKDLGADYVILGHSERRHYFGETDELVAKKVAHALEHGLKVIACIGETLEEREAGKTEEVVFRQTKALLAGLGDEWKN
2YPI: VVVAYEPVWAIGTGLAATPEDAQDIHASIRKFLASKLGDKAASELRILYGGSANGSNAVTFKDKADVDGFLVGGASLKPE
dTIM: VVIAYEPVWAIGTGKTATPEQAQEVHAFIRKWLAENVSAEVAESVRILYGGSVKPANAKELAAQPDIDGFLVGGASLKPE
2YPI: FVDIINSRN
dTIM: FLDIINSRN
```

O alinhamento acima reportou 102 mutações entre as proteínas 2YPI e dTIM. Levando em

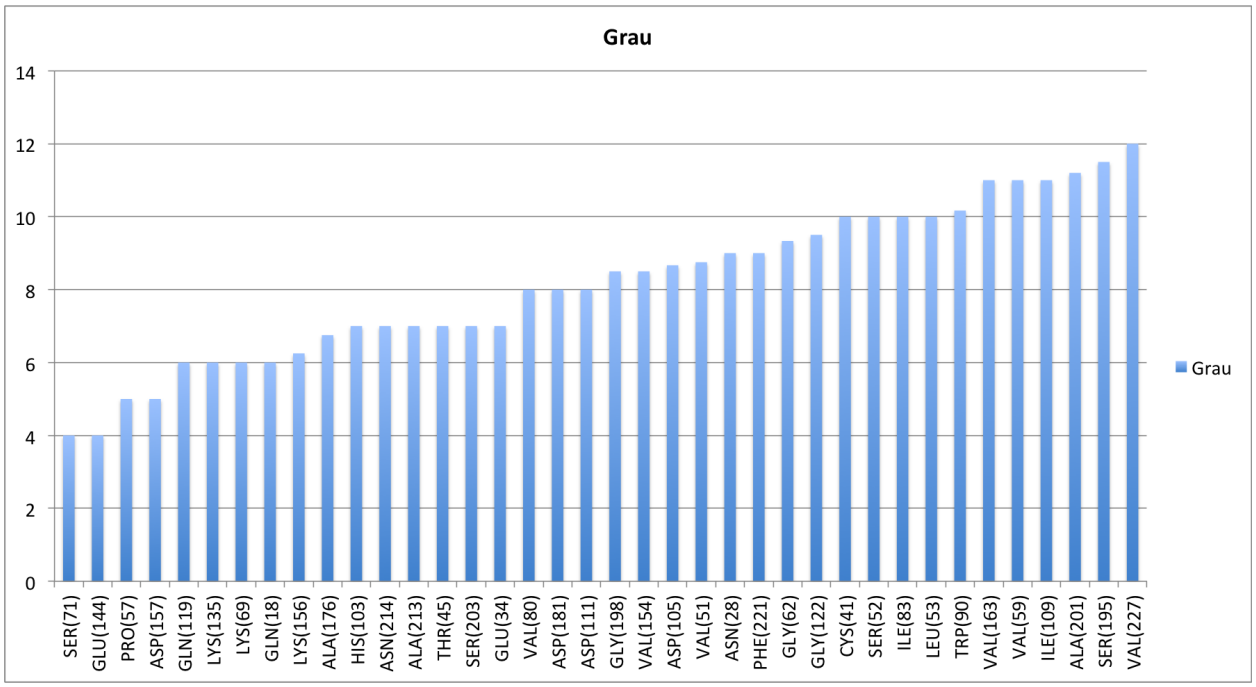
conta que o objetivo do presente trabalho seria o de escolher as dez mais influentes na função da proteína, julgou-se que é um número razoável: precisamos selecionar dez por cento destas como mutações impactantes.

Agora, com os dados do alinhamento e os grafos de toda a família obtidos, decidiu-se empregar um método simples para a análise da família. Como não podemos assumir que quaisquer dessas mutações ocorrerão nas mesmas posições nas outras proteínas da família, adotou-se como aproximação arrazoada a suposição de que se um determinado resíduo encontra-se na mesma posição em uma das proteínas da família, trata-se de um resíduo correspondente. Posto isso, buscamos então em todas as proteínas da família fornecidas, quais possuem quaisquer dos resíduos encontrados como mutados no alinhamento, nas mesmas posições em que se encontram na 2YPI. Calculamos as propriedades de grafo dos resíduos encontrados e, a título de análise, calculamos a média dessas propriedades em todas as ocorrências.

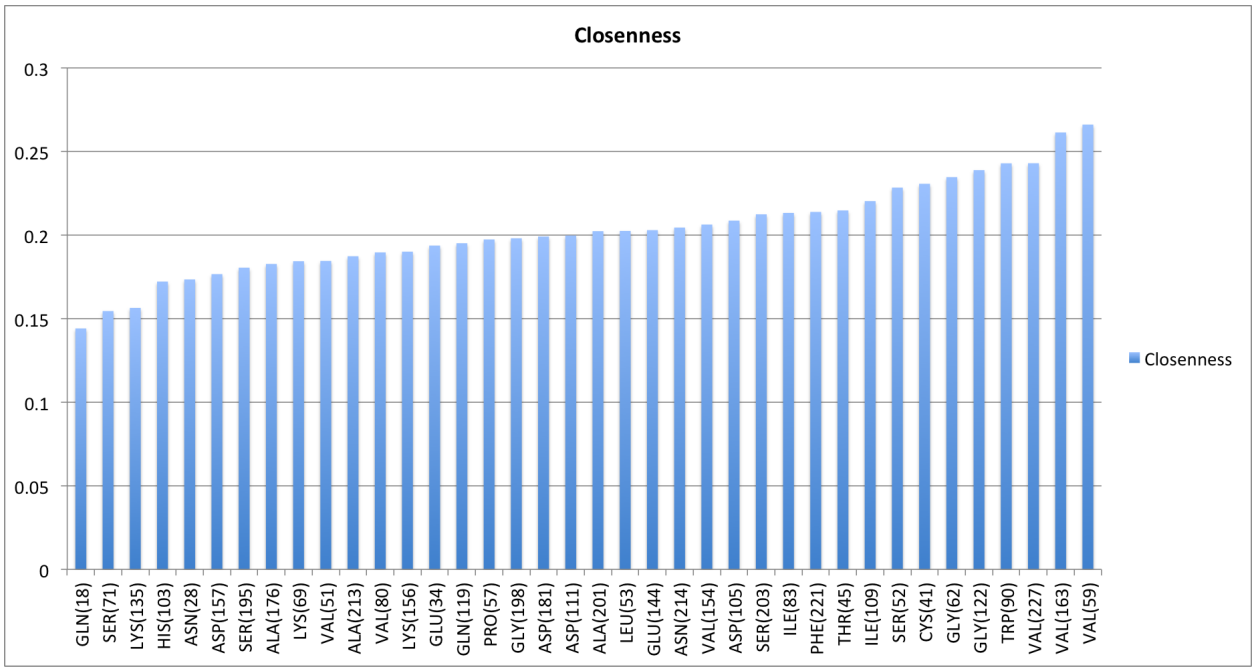
Um exemplo: nosso alinhamento encontrou uma mutação no resíduo ASP(157) da proteína 2YPI, que foi substituído por outro na dTIM. Pesquisamos então em todos os grafos de todas as proteínas da família, quais possuem um resíduo ASP na posição 157. Calculamos então as propriedades Grau, Closenness e Betweenness deste resíduo em toda a família. Caso ele ocorra em mais de uma proteína nessa mesma posição, os valores dessas propriedades serão acumulados e uma média aritmética simples será calculada.

Baseando-se na metodologia acima, filtramos 38 resíduos considerados mais relevantes dentre as 102 mutações encontradas. Os resultados da análise destes 38 resíduos podem ser avaliados nos gráficos que seguem.

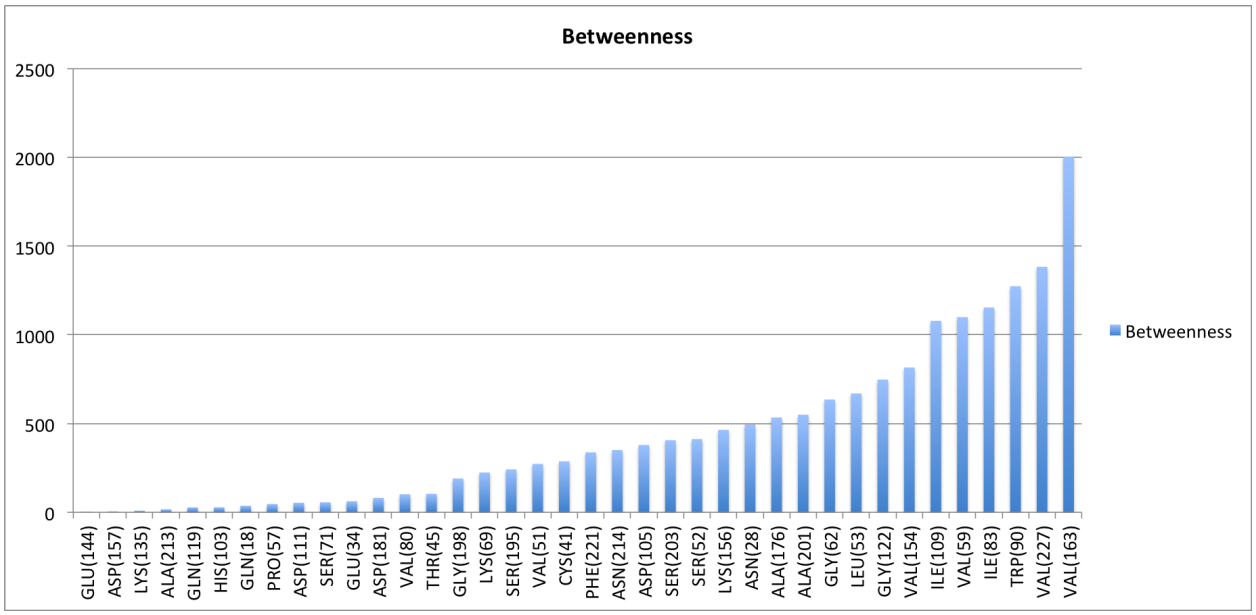
Grau


Grau
14
12
10
8
6
4
2
0
SER(71)
GLU(144)
PRO(57)
ASP(157)
GLN(119)
LYS(135)
LYS(69)
GLN(18)
LYS(156)
ALA(176)
HIS(103)
ASN(214)
ALA(213)
THR(45)
SER(203)
GLU(34)
VAL(80)
ASP(181)
ASP(111)
GLY(198)
VAL(154)
ASP(105)
VAL(51)
ASN(28)
PHE(221)
GLY(62)
GLY(122)
CYS(41)
SER(52)
ILE(83)
LEU(53)
TRP(90)
VAL(163)
VAL(59)
ILE(109)
ALA(201)
SER(195)
VAL(227)
Grau

Closenness


Closenness
0.3
0.25
0.2
0.15
0.1
0.05
0
GLN(18)
SER(71)
LYS(135)
HIS(103)
ASN(28)
ASP(157)
SER(195)
ALA(176)
LYS(69)
VAL(51)
ALA(213)
VAL(80)
LYS(156)
GLU(34)
GLN(119)
PRO(57)
GLY(198)
ASP(181)
ASP(111)
ALA(201)
LEU(53)
GLU(144)
ASN(214)
VAL(154)
ASP(105)
SER(203)
ILE(83)
PHE(221)
THR(45)
ILE(109)
SER(52)
CYS(41)
GLY(62)
GLY(122)
TRP(90)
VAL(227)
VAL(163)
VAL(59)
Closenness

Betweenness



Por se tratarem de muitos resíduos, uma análise visual não é suficiente para determinar a relevância dessas mutações. Por se tratarem, além disso, de três variáveis distintas, é ainda mais complicado fazer essas análises de forma leviana.

O método mais indicado, neste caso, seria utilizar um algoritmo de otimização para encontrar os máximos considerando todas as três variáveis. Por restrições de tempo, porém, optou-se por uma abordagem simplificada. Decidimos arbitrariamente por uma ordem de precedência dessas variáveis, no caso Grau -> Closenness -> Betweenness, e, usando um algoritmo de ordenação conservativa, ordenamos os resíduos encontrados sucessivamente por cada uma das variáveis. Ao fim do procedimento, selecionamos os dez primeiros resíduos como mais relevantes. O resultado dessa análise está disponível na conclusão do trabalho.

3.2 Estudo mais detalhado das mutações encontradas no TP1

Nesta seção, vamos pegar as mutações que encontramos no alinhamento do

trabalho prático 1, e verificar se aqueles resíduos da 2YPIA selvagem, que sofreram mutação, apresentam propriedades interessantes no grafo de contatos. Iremos dessa forma, analisar se um determinado resíduo que sofreu mutação tem grau alto, i.e. faz muito contatos, ou se ele é uma grafo central, calculado pelo *betweenness*.

Geramos o grafo para a proteína 2YPIA, em que os vértices são os resíduos e as arestas os contatos. O grafo tem 247 vértices e 991 arestas, e a explicação por ter um número menor de vértices que resíduos na sequência, é que no arquivo PDB não contém o primeiro resíduo.

Para cada uma das mutações geradas pelo alinhamento do TP1, calculamos o grau e o *betweenness* do resíduo referente a proteína 2YPIA, e a seguir estão duas listas com os resíduos que obtiveram melhores valores para cada uma dessas métricas. Na tabela temos o resíduo da sequência original, o resíduo que substituiu esse resíduo e a posição do resíduo da 2YPIA na sequência.

Tabela 1 - Grau

#	Posição Resíduo	Resíduo 2YPIA	Resíduo dTIM	Grau
1	11	F	W	12
2	90	W	Y	11
3	41	C	A	11
4	226	V	I	11
5	125	L	A	11
6	109	I	V	11
7	89	K	D	10
8	62	G	A	10
9	40	I	V	10
10	162	V	I	10

Tabela 2 - Betweenness

#	Posição Resíduo	Resíduo 2YPIA	Resíduo dTIM	Betweenness
1	204	L	V	2076
2	90	W	Y	1675
3	66	A	C	1646
4	62	G	A	1369
5	212	A	V	1368
6	109	I	V	1271
7	23	I	L	1194
8	41	C	A	1166
9	51	V	R	1160
10	226	V	I	1116

Tabela 3 - Closeness

#	Posição Resíduo	Resíduo 2YPIA	Resíduo dTIM	Closeness
1	195	K	N	0.0001084704
2	194	L	V	0.0001119028
3	18	Q	A	0.000115525
4	135	K	E	0.0001163809
5	191	F	W	0.0001180109
6	71	S	K	0.0001183235
7	196	L	V	0.0001186393
8	197	G	S	0.000118709
9	19	S	E	0.0001204568
10	29	T	A	0.0001206932

Das mutações selecionadas no TP1 como mais relevantes, apenas uma delas apareceu nas tabelas referentes aos resíduos com maior grau e maior betweenness. Essa mutação é referente a troca de um triptofano por uma tirosina (W por Y). Como podemos ver nas tabelas 1 e 2, o triptofano (90) possui o segundo maior grau e betweenness. Portanto, acreditamos que essa mutação realmente é impactante pra função, uma vez que ela foi consistente para as duas abordagens: alinhamento de sequências e estudo dos contatos. Um outro ponto que torna essa mutação ainda mais relevante, é o fato de envolver a troca de um triptofano, que é um dos aminoácidos que menos sofrem mutação.

Outras mutações que provavelmente tem alguma relação com a perda da função são aquelas que envolvem resíduos que têm o grau e o betweenness altos, pois isso significa que esse resíduos fazem muitos contatos e são resíduos centrais. Por isso, aquelas mutações que envolverem os resíduos que estão entre os dez com maior grau e betweenness serão consideradas como mais impactantes. São elas: (62)G->A, (41)C->A, (109)I->V, (226)V->I.

O impacto de trocar um resíduo que faz muitos contatos deve ser muito relevante, por isso vamos acrescentar à nossa lista a mutação (11)F->W, pois essa fenilalanina realiza 12 contatos, e poucos resíduos realiza essa quantidade de contatos, como podemos ver na figura 1.

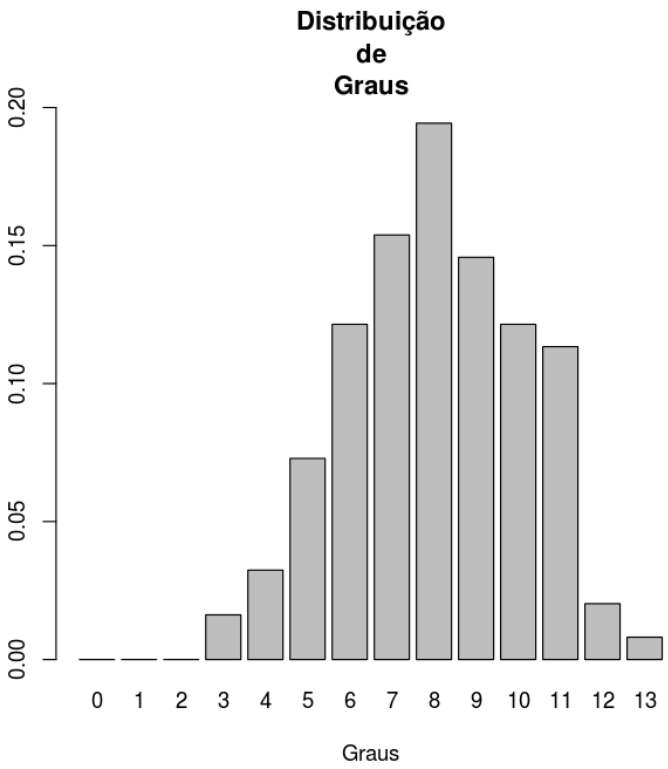


Figura 1 - Distribuição de Graus

4. CONCLUSÃO

Resíduos Considerados Mais Impactantes

Abaixo encontram-se tabulados os resíduos mutados considerados mais relevantes de acordo com ambas as abordagens implementadas. Nota-se uma correspondência de vários destes, indicando que ambas as abordagens possuem potencial, e que, provavelmente, as mutações destes resíduos, ocorridas entre dTIM e 2YPI, são especialmente relevantes.

Através da Análise de Família	**Através do Aprofundamento do TP1**
VAL(163)	CYS(41)
VAL(227)	PHE(11)
TRP(90)	TRP(90)
ILE(83)	VAL(226)
VAL(59)	LEU(125)
ILE(109)	ILE(109)
VAL(154)	LYS(89)
GLY(122)	-
LEU(53)	-
GLY(62)	GLY(62)

Foi interessante realizar o mesmo trabalho proposto no TP1, mas por uma perspectiva diferente, utilizando informações estruturais sobre as proteínas. Apenas uma das mutações foi considerada relevante nos dois trabalhos. Porém, essa mutação é especialmente interessante, pois envolve a troca de um triptofano, que é uma troca rara.

Apêndice 01: Dados dos Gráficos da Família

Resíduo	Betweenness	Closenness	Grau
ALA(176)	534.2419261	0.182673719	6.75
ALA(201)	549.9071842	0.202270371	11.2
ALA(213)	16.92611809	0.18729097	7
ASN(214)	350.6096474	0.204433498	7
ASN(28)	494.5585203	0.17341342	9
ASP(105)	379.281041	0.208602384	8.666666667
ASP(111)	53.87989845	0.19967923	8
ASP(157)	5.061909529	0.1765961	5
ASP(181)	81.33199742	0.199096135	8
CYS(41)	287.0360207	0.230628801	10
GLN(119)	27.43779372	0.195064899	6
GLN(18)	36.72490877	0.144097222	6
GLU(144)	1.669824053	0.202898551	4
GLU(34)	62.20039146	0.193666848	7
GLY(122)	747.7033467	0.238792196	9.5
GLY(198)	190.5255985	0.198030669	8.5
GLY(62)	635.5867991	0.234613522	9.333333333
HIS(103)	28.13672128	0.172145229	7
ILE(109)	1078.898387	0.220268113	11
ILE(83)	1153.978232	0.213218211	10
LEU(53)	669.6191731	0.202439024	10
LYS(135)	9.31471314	0.156407035	6
LYS(156)	464.5059988	0.190024112	6.25
LYS(69)	224.2906336	0.18430792	6
PHE(221)	337.6502843	0.213796058	9
PRO(57)	46.66699919	0.197325132	5
SER(195)	241.8545478	0.180447777	11.5
SER(203)	406.320532	0.212373842	7
SER(52)	412.6080805	0.22833843	10
SER(71)	56.26363028	0.15450908	4
THR(45)	104.1636071	0.214698389	7
TRP(90)	1273.731038	0.242872225	10.16666667
VAL(154)	815.9594384	0.206246515	8.5
VAL(163)	2004.424093	0.261319984	11
VAL(227)	1383.120453	0.242926829	12
VAL(51)	272.6213301	0.184486811	8.75
VAL(59)	1099.704571	0.266033254	11
VAL(80)	101.4690418	0.189589486	8