SANDRO RENATO DIAS

Tese de Doutorado

RESIDUE INTERACTION DATABASE - PROPOSIÇÃO DE MUTAÇÕES SÍTIO DIRIGIDAS COM BASE EM INTERAÇÕES OBSERVADAS EM PROTEÍNAS DE ESTRUTURA TRIDIMENSIONAL CONHECIDA

Tese apresentada ao Curso de Doutorado em Bioinformática, do Programa de Pós-Graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Bioinformática, Área de concentração: Estrutura de proteínas.

Orientador: Prof. Dr. Ronaldo Alves Pinto Nagem
Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, UFMG

Co-Orientador: Prof. Dr. Richard Charles Garrat
Departamento de Física e Informática, São Carlos, USP

Belo Horizonte – MG

Instituto de Ciências Biológicas da UFMG

2012

Dedico este trabalho à minha mãe, Laurita Veiga Dias, minha primeira professora, minha primeira orientadora, minha primeira mestra, a razão da minha existência. Foi à beira do tanque, enquanto lavava roupas, que me deu as minhas primeiras lições de alfabetização antes mesmo de eu iniciar os meus estudos na escola. Naquela época, por volta dos 5 ou 6 anos de idade, eu queria ler e ela me ensinou "Li-a be-be lei-te". D. Laurita me deu esperança, fé, confiança, me fez ser o que sou hoje e me dizia sempre "um dia, você vai ser doutor".

AGRADECIMENTOS

Apesar de ser uma seção de agradecimentos, vejo a necessidade de iniciá-la com um pedido de desculpas. Desculpas a todos que me rodearam pelas minhas falhas, ausências, atrasos, sonolências e demais consequências dessa vida de doutorando. Em se tratando de agradecimentos, primeiramente a Deus, pela luz, força, persistência e aquele conforto no momento de maior desespero. A meus pais, sempre, incondicionalmente. Ao meu filho Gabriel pelo carinho, compreensão, apoio e por me deixar vencer no Mortal Kombat, mesmo eu estando dormindo. À minha Débora^{BBGBBST} *por tudo e principalmente por me aturar e estar ao meu lado, me apoiando. Ao meu orientador e amigo Prof. Dr. Ronaldo Alves Pinto Nagem, por me permitir usufruir da sua sabedoria através dos seus ensinamentos e observações precisas. Aos colegas do laboratório Bioest, aos meus amigos, familiares e a todos que colaboraram direta ou indiretamente com este trabalho. Por fim, um agradecimento especial ao meu fiel antigo companheiro notebook Toshiba e meu atual companheiro notebook Dell, amigos que me acompanharam durante todo o doutorado colaborando para o progresso da Ciência. Eles estiveram mais próximos de mim e tiveram mais contato comigo do que qualquer outra pessoa.*

"Tudo vale a pena se a alma não é pequena" porque *"o homem é do tamanho do seu sonho"*. Assim, *"eu sei que não sou nada e que talvez nunca tenha tudo. Aparte isso, eu tenho em mim todos os sonhos do mundo."* E aí eu me pergunto, depois de tudo: *"Valeu a pena? Tudo vale a pena se a alma não é pequena. Quem quer passar além do Bojador tem que passar além da dor. Deus ao mar o perigo e o abismo deu, mas nele é que espelhou o céu."*

Fernando Pessoa (vários trechos intercalados)

RESUMO

Neste trabalho é descrito um algoritmo usado para predizer pares de resíduos de aminoácidos em proteínas alvo (com estrutura tridimensional conhecida) que poderiam ser mutados por pares diferentes de resíduos de aminoácidos com o objetivo de introduzir uma nova/diferente interação entre estes resíduos. Isto resulta em um mutante "in silico" com possibilidades estereoquímicas de existir "in vitro" com o aumento da estabilidade conformacional e térmica. Para alcançar isso, foi criado um banco de dados baseado no PDB composto de pares de resíduos de aminoácidos interagentes observados em proteínas de estrutura conhecida. As mutações são propostas de forma a manter o enovelamento da proteína alvo (e consequentemente sua função) através, basicamente, da conservação da conformação da cadeia principal dos resíduos mutados. Neste trabalho também são apresentados os aspectos principais dessa base de dados, a forma como encontrar os pontos de mutação e alguns resultados. Uma busca completa na estrutura de uma proteína alvo foi realizada para identificar cada par que poderia ser mutado usando alguns dos pares do banco. Pretende-se com este procedimento, verificar um número de possíveis mutantes em diferentes enzimas com potencial de aplicação em processos de biorremediação, onde condições ambientais agressivas são esperadas. É apresentada a ferramenta RID (*Residue Interaction Database*), uma nova base de dados e algoritmo para propor mutação de pares de resíduos em uma proteína objetivando aumentar sua estabilidade focando na manutenção da conformação de sua cadeia principal. São descritos os detalhes do algoritmo para gerar a base de dados dos resíduos de aminoácidos interagentes e o método para otimizar a busca. Comparado com outros métodos, RID aumenta as alternativas de proposição de mutação devido à variedade de interações usadas para criar o banco de dados e que irão contribuir para o aumento da estabilidade proteica. A ferramenta se encontra disponível em http://www.bioest.icb.ufmg.br/RID.

Palavras-chave: interação resíduo-resíduo, banco de dados biológico, modificação de proteína, mutação sítio dirigida.

ABSTRACT

In this work we describe an algorithm which is used to predict amino acid residue pairs in target proteins (with known 3D structure) that could be replaced by a different amino acid residue pair in order to introduce a new/different interaction between residues. This might result in an “in silico” mutant with stereochemistry possibilities to exist “in vitro” with increased thermo and conformational stability. To address this, we have created a PDB-based database composed of pairs of interacting amino acid residues observed in proteins with known structure. The mutations are proposed in a way to maintain the target protein's fold (and function) as, basically, the main chain conformation of mutated residues are supposed to be conserved. In this work we also present the main aspects of this database, the way to find the mutation points and some results. A complete search in a target protein structure was performed to identify each residue pair that could be mutated using some of the pairs in the database. We intend to use this procedure to verify a number of possible mutations in different enzymes with potential application in bioremediation processes, where aggressive environmental conditions are expected. We present RID (Residue Interaction Database), a novel database and algorithm to propose a residue pair mutation in a protein aiming to increase its stability focusing in the conformation maintainability. We describe the details of the algorithm do generate the database of interacting amino acid residues and the method to optimize the database for quick searches. Compared to other methods, RID increases the alternatives to propose mutation because of the variety of the interactions used to create the database and that will contribute to increase the protein stability. The tool is available at: http://www.bioest.icb.ufmg.br/RID.

Keywords: residue-residue interaction, database, protein modification, direct site mutagenesis.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE EQUAÇÕES

LISTA DE ALGORITMOS

LISTA DE TABELAS

SUMÁRIO

1. Introdução

As proteínas constituem uma das classes mais importantes de macromoléculas existentes nos seres vivos devido às inúmeras funções que desempenham, indo desde catálise, transporte, regulação até função imune (Lehninger, Nelson, Cox, 2007). Pesquisas que envolvam, portanto, melhorias nas estruturas dessas macromoléculas beneficiam vários segmentos distintos como saúde (desenvolvimento de fármacos, tratamentos, vacinas), indústria (aprimoramento em enzimas digestivas usadas em vários processos), meio-ambiente (alterações em enzimas para a degradação de contaminantes), dentre inúmeros outros. Segundo Teilum, Olsen e Kragelund (2011), estas melhorias podem envolver a mutação de um ou mais aminoácidos, que são os blocos mínimos que compõe a proteína, visando um aumento ou diminuição de estabilidade ou flexibilidade com a manutenção ou não da função da proteína.

O estudo das mutações em proteínas tem aumentado o entendimento geral sobre as forças que estabilizam estas macromoléculas e também sobre a contribuição de cada uma dessas forças nas etapas de enovelamento e desenovelamento (Pace *et al*, 1996; Huang *et al*, 2007; González-Díaz, Molina e Uriarte, 2005; Gromiha, 2010). A estrutura tridimensional de uma proteína, ou sua conformação, está diretamente relacionada à sua função, o que significa que qualquer alteração em sua estrutura nativa poderá afetar seu mecanismo de ação no organismo. Podem-se citar, como exemplo, as doenças provocadas por mutações gênicas, que alteram a sequência de aminoácidos da proteína que o gene codifica, como fibrose cística, daltonismo e hemofilia (Voet, Voet, 2011; Motta, 2011; Lehninger, Nelson, Cox, 2007).

A estabilidade conformacional de uma proteína pode ser definida como a variação da energia livre de Gibbs (ΔG, Equação 1), para a reação de enovelamento $\leftrightarrow$ desenovelamento sob condições fisiológicas (Pace *et al*, 1996; Magliery, Lavinder e Sullivan, 2011). Já em outro trabalho, Pace (1995) define $\Delta(\Delta G)$, Equação 2, como uma medida para a comparação de

proteínas mutantes (mut) e proteínas nativas (wt) em termos da diferença da estabilidade conformacional.

$$\Delta G = \Delta H - T\Delta S$$

Equação 1 – Variação da Energia Livre de Gibbs[1]

$$\Delta(\Delta G) = \Delta G(wt) - \Delta G(mut)$$

Equação 2 – Medida para comparação em termos da diferença da estabilidade conformacional

Diante do fruto das pesquisas relacionadas à estabilidade, estrutura e função de proteínas, tem-se como foco desse projeto de pesquisa a seguinte formulação: dada uma proteína de estrutura tridimensional conhecida, quais possíveis pares de aminoácidos poderiam ser mutados concomitantemente para que uma nova interação pudesse se formar a partir desses novos resíduos de aminoácidos viabilizando uma proteína mais estável? Além disto, como fazer para que estas mutações não interfiram com a funcionalidade da proteína e seu enovelamento? No nosso entendimento estas perguntas serão respondidas por meio de uma análise das estruturas tridimensionais de proteínas conhecidas e a formulação de um algoritmo de busca de padrões de conformações das cadeias principais dos pares de resíduos de aminoácidos interagentes.

1.1 Proteínas

Uma proteína é uma macromolécula, das mais abundantes nos seres vivos, constituída a partir de um conjunto ubíquo de 20 aminoácidos distintos ligados covalentemente em sequências lineares (Lehninger, Nelson e Cox, 2007). Elas servem para funções cruciais em essencialmente todos os processos biológicos, como catalisadores, transporte, armazenamento, apoio mecânico, proteção imunitária, regulação, estruturais, diferenciação celular, geração de movimento, transmissão de impulsos nervosos, dentre outras (Berg, Tymoczko, Stryer, 2006). Motta (2011) acrescenta à esta lista "a manutenção da distribuição de água entre o compartimento intersticial e o sistema vascular do organismo, participação na homeostase e coagulação sanguínea, nutrição de tecidos, formação de tampões para manutenção de pH, etc".

[1] onde ΔH é a entalpia, T a temperatura (em Kelvin) e ΔS a entropia envolvidas (Lehninger, Nelson, Cox, 2007).

1.1.1 Aminoácidos

Aminoácidos constituem a estrutura básica de uma proteína. Os aminoácidos são compostos por uma cadeia principal de átomos (N, Cα, C, O), que efetua a ligação peptídica com a cadeia principal de outro aminoácido, e uma cadeia lateral R (ligada ao Cα, iniciando-se no Cβ, exceto para a Glicina, que não possui), onde reside a particularidade do aminoácido (Richardson, 1981).

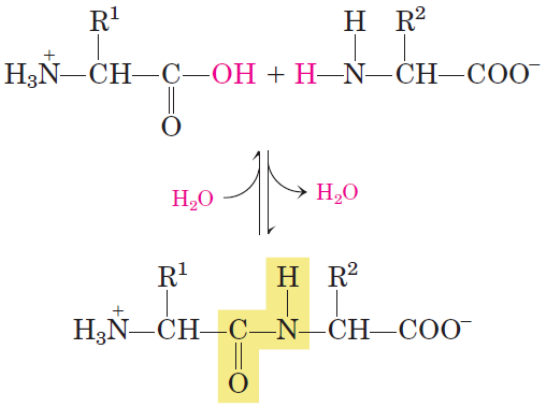


Figura 1 - Formação da ligação peptídica por condensação (Lehninger, Nelson e Cox, 2007)

Laskowski, Moss e Thornton (1993) estudaram os melhores valores (identificados na Figura 2 - esquerda) para as 5 distâncias atômicas que envolvem estes átomos da cadeia principal (N-Cα, Cα-Cβ, Cα-C, C-O, C-N) e 7 ângulos relacionados (C-N-Cα, N-Cα-Cβ, N-Cα-C, Cβ-Cα-C, Cα-C-N, Cα-C-O, O-C-N) da cadeia principal das proteínas. Na Figura 2, à direita, é possível observar os valores dessas distâncias e ângulos definidos por Voet e Voet (2011). Laskowski, Moss e Thornton (1993) demonstraram que estes dados além de precisos devem ser consistentes durante um processo experimental (ou teórico) de determinação estrutural, uma vez que a estrutura está diretamente relacionada com a funcionalidade e qualquer erro cometido neste processo pode ser determinante para a classificação do modelo 3D gerado.

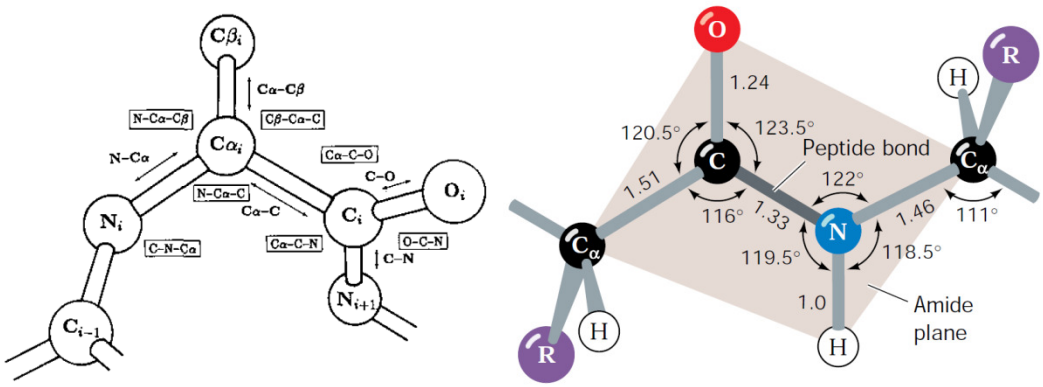


Figura 2 - Distâncias e ângulos da cadeia principal. Estudados por Laskowski, Moss e Thornton (1993) – à esquerda. Apresentados por Voet e Voet (2011) – à direita.

A Figura 3 apresenta um diagrama de Venn agrupando os aminoácidos por categorias e indica, além do nome, o mnemônico (três letras) e o símbolo (uma letra) associado a este aminoácido, bem como a estrutura de suas cadeias principal e lateral. Estas categorias referem-se à cadeia lateral e estão diretamente ligadas à estrutura e ao enovelamento protéico. Lehninger, Nelson e Cox (2007) afirmam que cada proteína tem uma função estrutural e química específica, diretamente relacionada com sua estrutura tridimensional.

As categorias apresentadas na Figura 3 são definidas por Brevern (2006), Livingstone e Barton (1993), assim como por Taylor (1986):

- Pequenos – classificação diretamente relacionada ao tamanho da cadeia lateral, e o volume que ocupa, estando nesta categoria apenas os aminoácidos com volume inferior a 60 $Å^3$.

- Curtos – uma subcategoria da anterior, onde se encontram os aminoácidos com cadeia de até três átomos (não H) e com volume inferior a 35 $Å^3$. A Cisteína faz parte desse grupo devido ao estado de oxidação da ligação S-H polarizada sugerindo similaridade à Serina (O-H). Já a Cistina faz parte da categoria anterior pois, apesar do tamanho curto da cadeia (2 átomos não H), a formação de ponte dissulfeto implica em um aumento significativo do volume.

- Polares – seus grupos R são mais solúveis em água (hidrofílicos) pois contém grupos funcionais que tendem a participar de ligações de hidrogênio com a água.

- Hidrofóbicos – que possuem menor afinidade pelo solvente polar – água –, voltando-se para o interior da proteína.

- Aromáticos – que possuem anel aromático em sua composição, com suas cadeias laterais relativamente apolares participam de interações hidrofóbicas.

- Prolina é um aminoácido que possui propensão a conectar-se tanto a resíduos hidrofílicos quanto hidrofóbicos, o que o impede de ser classificado em qualquer uma das duas categorias.

- Carregados – que são energeticamente favoráveis a reações com a água, também definidos como aqueles que se apresentam normalmente completamente ionizados, positiva e negativamente.

- Alifáticos – não polares e hidrofóbicos, possuindo apenas carbono e hidrogênio em sua formação, além de não conterem anéis aromáticos.

Betts e Russell (2003) apresentam um diagrama similar online[2] com a possibilidade de navegar nas características de cada grupo além das características individuais de cada resíduo. Também afirmam que esta é uma das muitas possíveis classificações e que é uma das que cobre o maior número de categorias e, portanto, mais aceita na Literatura.

[2] Betts M.J., Russell R.B. Amino acid properties and consequences of subsitutions. In *Bioinformatics for Geneticists*, M.R. Barnes, I.C. Gray eds, Wiley, 2003. Disponível em: <http://www.russelllab.org/aas/>. Acessado em: 20/09/2012.

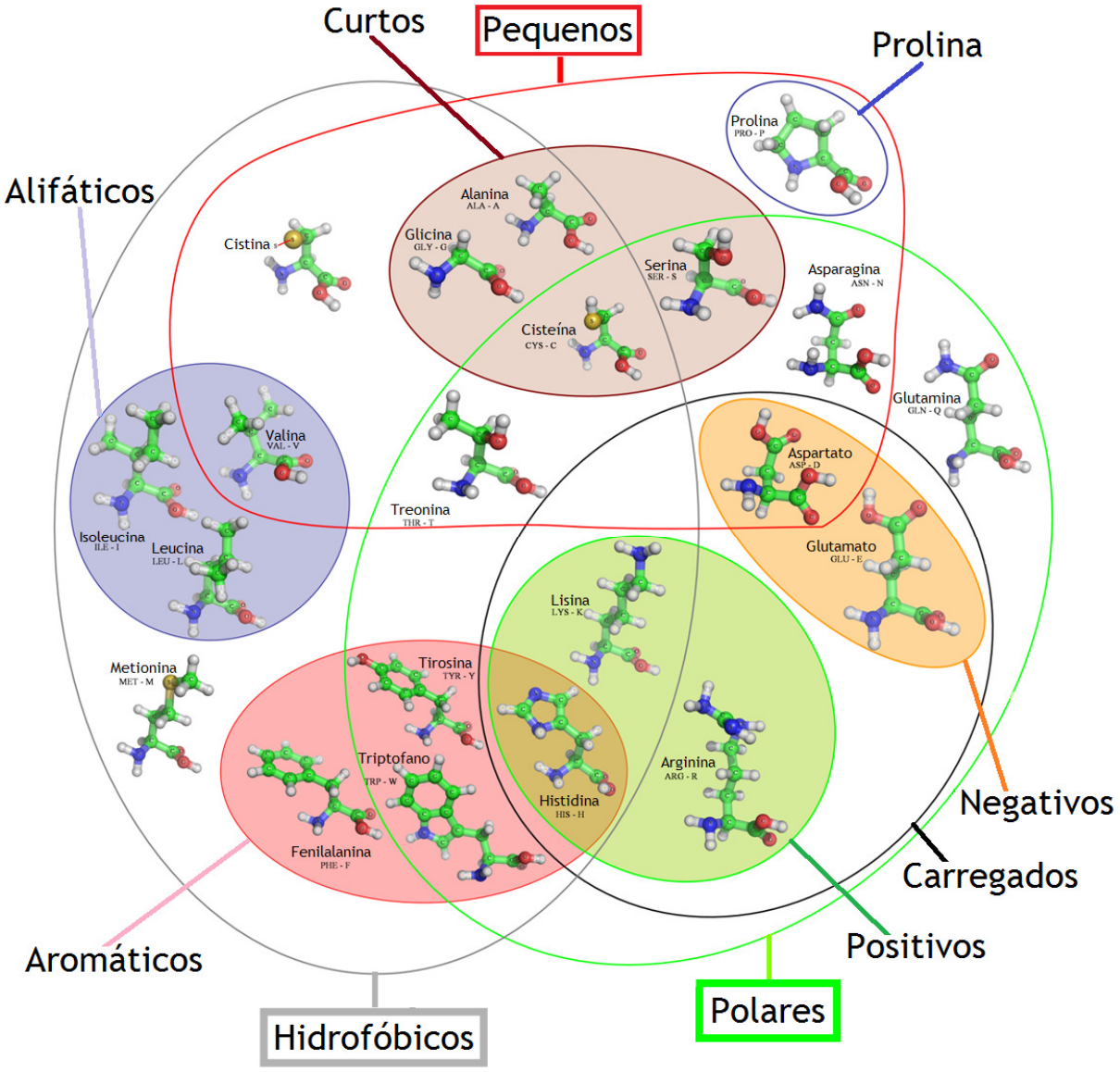


Figura 3 – Aminoácidos agrupados em categorias num Diagrama de Venn

Fonte: Adaptado de Brevern (2006), que foi baseado em Livingstone e Barton (1993)

1.1.2 Estrutura

Segundo Lehninger, Nelson e Cox (2007), a estrutura primária de uma proteína refere-se à sequência de resíduos de aminoácidos que a compõem. Um polipeptídeo, formado pelas ligações peptídicas entre aminoácidos da estrutura primária, se enovela formando a estrutura terciária através de elementos de estrutura secundária como, por exemplo, as α-hélices (estruturas helicoidais em azul na Figura 4), fitas β (setas vermelhas na Figura 4, que agrupadas formam as folhas β) e regiões de loop ou alças (estruturas em ciano, amarelo e

verde na Figura 4). As proteínas com mais de uma cadeia polipeptídica apresentam por fim uma estrutura quaternária, como no exemplo da Figura 5.

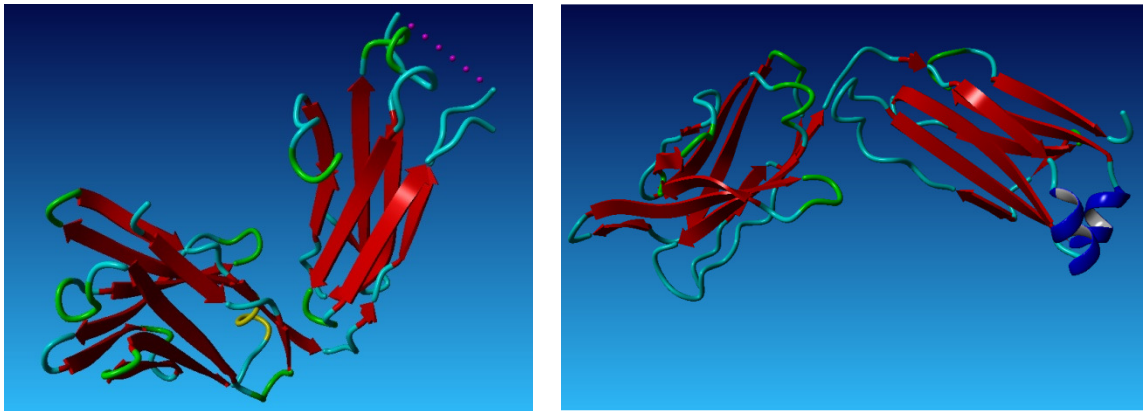


Figura 4 - Estruturas terciárias da proteína 1BBD[3] (PDB). À esquerda, cadeia L da proteína. À direita, cadeia H da mesma proteína.

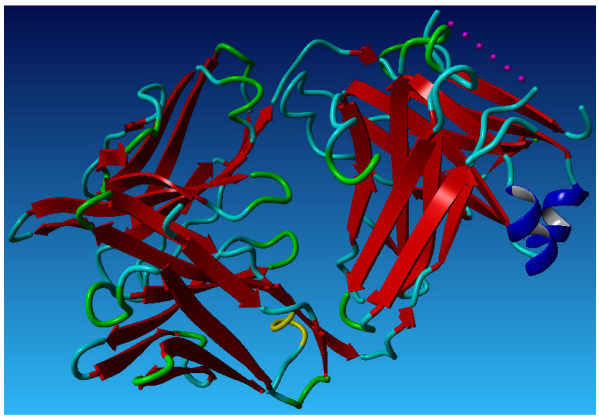


Figura 5 - Estrutura quaternária da proteína 1BBD (PDB), agrupando as duas estruturas terciárias da Figura 4.

Tradicionalmente representa-se a estrutura primária de uma proteína pela sequência de seus aminoácidos constituintes em ordem de ligação, conforme mostrado na Figura 6. Isto é feito representando cada resíduo de aminoácido pelo seu mnemônico de uma letra na ordem que eles ocorrem na proteína.

[3] PDB ID: 1BBD. Tormo, J.; Stadler, E.; Skern, T.; Auer, H.; Kanzler, O.; Betzel, C.; Blaas, D.; Fita, I. Three-dimensional structure of the Fab fragment of a neutralizing antibody to human rhinovirus serotype 2. Journal: (1992) *Protein Sci.* 1: 1154-1161. PubMed: 1338980. PubMedCentral: PMC2142184. DOI: 10.1002/pro.5560010909.

```
>1BBD:L|PDBID|CHAIN|SEQUENCE
DIVMTQSPSSLTVTTGEKVTMTCKSSQSLLNSRTQKNYLTWYQQKPGQSPKLLIYWASTRESGV
PDRFTGSGSGTDFTLSISGVQAEDLAVYYCQNNYNYPLTFGAGTKLELKRADAAPTVSIFPPSS
EQLTSGGASVVCFLNNFYPKDINVKWKIDGSERQNGVLNSWTDQDSKDSTYSMSSTLTLTKDEY
ERHNSYTCEATHKTSTSPIVKSFNRNEC
```

Figura 6 - Sequência de resíduos de aminoácidos na proteína cuja estrutura tridimensional foi resolvida e depositada sob o código 1BBD no *Protein Data Bank* (PDB). Por uma questão de simplificação, será adotado o código PDB para se referir à proteína cuja estrutura foi determinada e depositada no PDB sob este mesmo código.

Gromiha e Selvaraj (2004), analisaram vários fatores que envolvem as interações inter-resíduos e a estabilidade de uma proteína. Segundo os autores, as interações hidrofóbicas são as forças dominantes mas as ligações de hidrogênio contribuem significantemente para a estabilidade do estado nativo. Durante o processo de enovelamento protéico, as forças hidrofóbicas direcionam a cadeia polipeptídica rumo ao estado enovelado superando os fatores entrópicos (considerando-se apenas o polipeptídeo), enquanto que as ligações de hidrogênio, pares iônicos, pontes dissulfeto e interações de van der Waals definem a forma e mantém a estrutura estável (Ponnuswamy e Gromiha, 1994, *apud* Gromiha e Selvaraj, 2004).

1.1.3 Interações e estabilidade

Segundo Lehninger, Nelson e Cox (2007) o termo estabilidade pode ser definido como a tendência em manter a conformação nativa da proteína. As proteínas nativas, segundo os autores, são apenas ligeiramente estáveis, pois a variação de energia entre os estados enovelado e desenovelado (ΔG) está na faixa de 20 a 65 kJ/mol, em condições fisiológicas. As ligações covalentes individuais, bem mais fortes (200 a 460 kJ/mol para serem quebradas) que as interações fracas individuais (4 a 30 kJ/mol para serem quebradas), contribuem pouco para a diferença energética entre os estados enovelados e desenovelados, já que não se alteram no processo. Por outro lado, as interações fracas, como as ligações de hidrogênio, apesar de contribuírem pouco individualmente se encontram em grande quantidade na estrutura (Lenhinger, Nelson e Cox, 2007; Voet e Voet, 2011).

Vários autores estudam as forças envolvidas no enovelamento e estabilidade das proteínas. Dill (1990) apresenta uma análise dessas forças considerando suas contribuições para o enovelamento dos polipeptídeos. Pace (1990) descreve métodos de medida da estabilidade

conformacional de proteínas globulares e discute abordagens usadas para aumentar sua estabilidade. Os métodos abordados por ele, dentre outros desenvolvidos para outros tipos de proteínas foram estudados também por Osherovich (2011), Gromiha (2010), Cohen, Potatov e Schreiber (2009) além de Franks (2002). Cohen, Potatov e Schreiber (2009) também implementaram métodos para predizer a estabilidade da proteína mediante algumas mutações propostas.

Fágáin (1995) em seu trabalho faz uma revisão sobre ensaios que permitiram observar a perda de função por algumas proteínas a partir de mutações induzidas nestas estruturas. O autor aborda ainda em seu trabalho vários tipos de interação molecular que puderam aumentar a estabilidade das proteínas a partir, também, de mutações induzidas. Em trabalho posterior, Fágáin (2011) incrementa o trabalho anterior definindo a estabilidade de uma forma precisa e contextualizada com uma lista de 8 métodos e índices para medir a estabilidade de uma proteína. Para o autor, a estabilidade refere-se a uma resistência da proteína a influências adversas como calor, por exemplo.

Pace *et al* (2011) definem que a estabilidade total de uma proteína é a soma da contribuição de pontes dissulfeto, ligações de hidrogênio e interações hidrofóbicas. Os autores também indicaram o percentual de contribuição de cada interação para a estabilidade da proteína. Considerando um grupo de 22 proteínas alvo específicas, a média de contribuição das interações hidrofóbicas para a estabilidade foi de 60±4%, a menor contribuição ocorreu na proteína RNase T1 PDB 9rnt (54%) e a maior na proteína barstar PDB 1bta (73%). Já as ligações de hidrogênio apresentaram uma contribuição média de 40±4%, a menor ocorreu na proteína barstar (27%) e a maior na proteína RNase T1 (43%). A maior contribuição das pontes dissulfeto para a estabilidade foi de 5%, na proteína RNase A PDB 9rsa. Apenas 8 das proteínas estudadas possuíam pontes dissulfeto. Hinz *et al* (1993) e Lins e Brasseur (1995) já haviam citado estas interações como responsáveis pela estabilidade de uma proteína e incluíram em seu estudo também a interação de van der Waals para compor este grupo. Lehninger, Nelson e Cox (2007) também defendem que pontes dissulfeto, assim como interações não covalentes fracas como ligações de hidrogênio, interações hidrofóbicas e iônicas são responsáveis pela estabilidade da conformação nativa da proteína. Estes ainda ressaltam o papel dessas interações fracas no enovelamento da cadeia polipeptídica em estruturas secundárias e terciárias específicas e seus agrupamentos com outros polipeptídeos para formar as estruturas quaternárias.

Em termos de interações específicas (pontes, ligações) há diversos trabalhos na Literatura que as descrevem e, principalmente, caracterizam com foco em estabilidade. Singh e Thornton (1992) apresentam um atlas de 400 interações possíveis par a par para os 20 diferentes tipos de aminoácidos. Em seu trabalho, clusterizam as possíveis interações medindo distâncias e ângulos entre os resíduos interagentes, apresentando-os em forma de gráficos, visualizações moleculares, clusteres e distribuições, disponíveis online[4].

A Figura 7 apresenta algumas das interações que afetam a estabilidade de uma proteína. Estas interações serão abordadas nas seções seguintes.

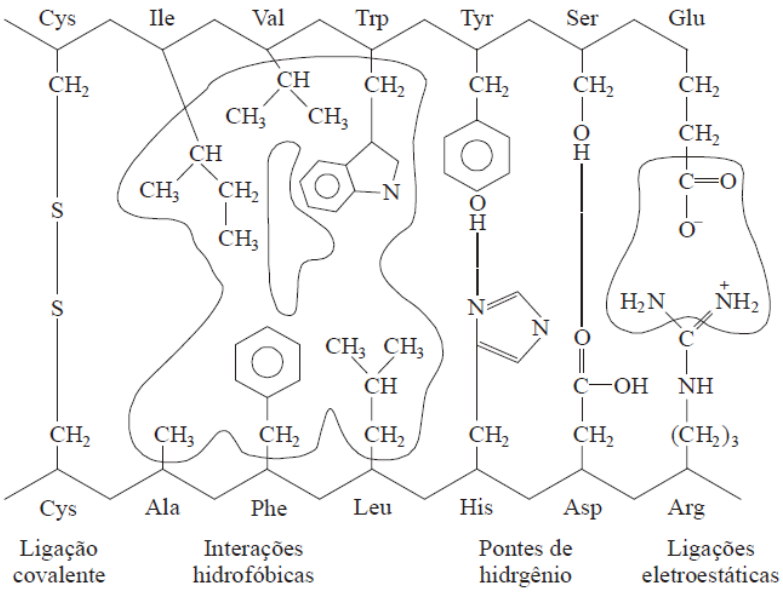


Figura 7 - Algumas interações que afetam a estabilidade de uma proteína

Fonte: Motta (2011)

1.1.3.1 Interações eletrostáticas

Interações eletrostáticas (como vistas na Figura 7) ocorrem entre partículas carregadas e a energia envolvida é dada, em função da distância entre as duas cargas puntuais próximas, pela lei de Coulomb (Equação 3), onde D_{ij} é a distância entre as cargas i e j, q_i e q_j são as respectivas frações das unidades de carga, ε é a constante dielétrica do solvente. O valor da constante dielétrica em interfaces lipídio/água gira em torno de 10 a 40, o mesmo se dará,

[4] Singh J., Thornton J.M. *Atlas of Protein Side-Chain Interactions, Vols. I & II*, IRL press, Oxford, 1992. Dataset disponível em: <http://www.biochem.ucl.ac.uk/bsm/sidechains>. Acessado em: 15/01/2011.

provavelmente, em proteínas, considerando a similaridade entre a estrutura da proteína e de uma membrana (Lins e Brasseur, 1995).

$$Eelec_{ij} = \frac{q_i q_j}{\epsilon D_{ij}}$$

Equação 3 – Energia Eletrostática

Voet e Voet (2011) afirmam que o cálculo do potencial eletrostático de uma proteína envolve sofisticados recursos matemáticos e computacionais como o programa GRASP[5] (*Graphical Representation and Analysis of Surface Properties*) usado para calcular o potencial da superfície eletrostática de uma proteína. As interações eletrostáticas podem ocorrer quando, por exemplo, grupos carregados positivamente como os grupos amino ($-NH_3^+$) das cadeias laterais das lisinas interagem com os grupos carboxila ($-COO^-$) carregados negativamente do ácido glutâmico ou ácido aspártico (Motta, 2011).

1.1.3.2 Pontes dissulfeto

As pontes dissulfeto são ligações covalentes formadas pela interação entre os átomos de enxofre das cisteínas, que após serem oxidadas se tornam cistinas. Estas pontes são responsáveis por manter a estabilidade conformacional (estrutura tridimensional) de uma proteína ligando partes distantes de uma cadeia polipeptídica ou cadeias diferentes. Por esta razão as cisteínas tem um papel especial na determinação da estrutura tridimensional das proteínas (Hunter, 1993). A Figura 8 apresenta uma ponte dissulfeto, os átomos SG (enxofre gama, átomos verdes da figura) de cada uma das cisteínas se ligam estabilizando as cadeias (trecho N-CA-C-O de cada lado).

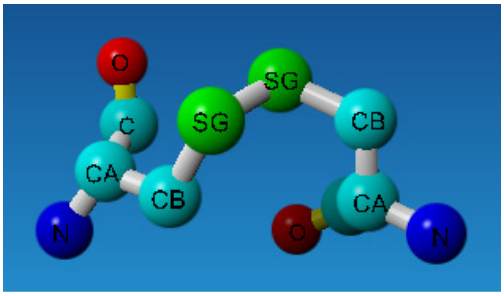


Figura 8 - Ponte dissulfeto ligando duas cisteínas (formando as cistinas)

[5] Disponível em: http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:GRASP. Acessado em: 20/07/2012.

Petersen, Jonson e Petersen (1999) descreveram alguns dados de interações de cisteínas e concluíram que nem toda ponte dissulfeto aumenta a estabilidade de uma proteína. Afirmaram ainda que em alguns casos, não houve aumento e nem diminuição da estabilidade das proteínas observadas, citando inclusive outros trabalhos que apresentam também estas conclusões (Wells e Powers, 1986; Matsumara *et al*, 1989; Betz, 1993). Segundo os autores, a estabilidade resultante de uma ponte dissulfeto é determinada pela geometria da ligação bem como pelas interações com o restante da proteína, definem também que é altamente relevante avaliar esta importância estrutural, uma vez que estes fatores apontam a possibilidade ou não de contribuição para a estabilidade protéica. No trabalho citado, foram avaliadas interações provenientes de proteínas de cadeias únicas somente (monômeros), contendo pelo menos uma ponte dissulfeto, num total de 131 proteínas. As pontes dissulfeto estudadas forneceram resultados e validações sobre sua importância para a estabilidade. Mason *et al* (2012) também confirmaram o papel da ponte dissulfeto na estabilidade e enovelamento da proteína, estudando o citocromo c_{6A} de *Arabidopsis thaliana*.

1.1.3.3 Ligações de Hidrogênio

As ligações de hidrogênio, segundo Rose *et al* (2006), Jackson (2005), Efting e Pedigo (2003), dominam o processo de enovelamento das proteínas, pois são responsáveis pela manutenção das estruturas secundárias como α-hélices e folhas β. Isto acontece, pois esta interação ocorre devido ao compartilhamento de um átomo de hidrogênio entre um doador de próton e um aceptor de próton (Figura 9), sendo interpretada como um estágio intermediário na transferência de um próton de um ácido AH para uma base B, ocorrendo entre moléculas polares (Lins e Brasseur, 1995).

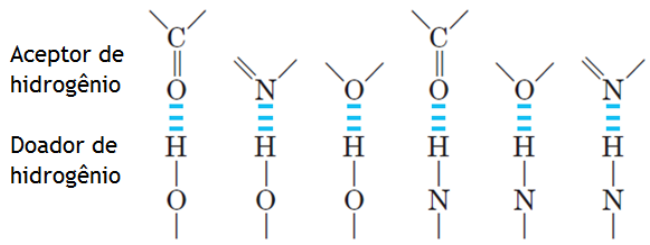


Figura 9 - Ligações de hidrogênio comuns em sistemas biológicos. Acima, aceptores de H e abaixo, doadores.
Fonte: Lehninger, Nelson, Cox (2007)

A Figura 10 apresenta duas ligações de hidrogênio: uma entre o átomo OD2 (oxigênio delta 2) do resíduo aspartato (que foi isolado do restante da cadeia) e o H (hidrogênio) da molécula de água próxima ao resíduo; a outra é entre os átomos OD1 (oxigênio delta 1) do resíduo e o H ligado ao N (nitrogênio) da cadeia principal do resíduo.

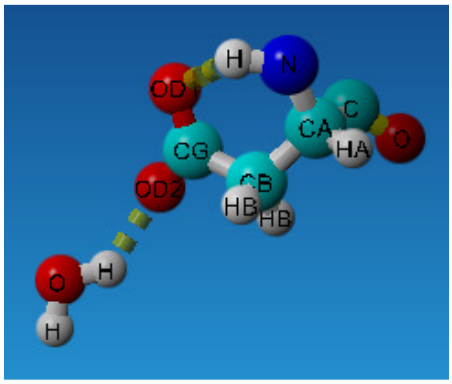


Figura 10 – 2 ligações de hidrogênio (OD2-H e OD1-H)

Em seus resultados disponíveis online[6], McDonald e Thornton (1993) demonstram a distribuição das frequências e geometrias das ligações de hidrogênio formadas por doadores e aceptores tanto da cadeia principal quanto da cadeia lateral dos resíduos, objetivando apresentar a probabilidade de uma ligação de hidrogênio em cada possível interação entre resíduos. Eswar e Ramakrishnan (2000) apresentam os motivos compostos de resíduos polares que apresentam probabilidade maior de ligações de hidrogênio, assim como a propensão de vários resíduos individuais a formar tal ligação, considerando os átomos envolvidos e o número de exemplos encontrados na base pesquisada, um conjunto de 250 proteínas não homólogas com alta resolução, extraídas do PDB.

1.1.3.4 Interações Hidrofóbicas

Voet e Voet (2011) definem o efeito hidrofóbico como o nome dado às influências que levam substâncias apolares a minimizar seus contatos entre água e moléculas anfifílicas[7], como sabões e detergentes, para formar micelas em soluções aquosas. Assim, interações hidrofóbicas são as forças que mantém as regiões apolares dessas moléculas juntas (Lehninger, Nelson, Cox, 2007). Estes ainda complementam afirmando que interações

[6] McDonald I., Thornton J.M. Atlas of Side-Chain and Main-Chain Hydrogen Bonding. Web edition 1994. Original edition 1993. Disponível em: <http://www.biochem.ucl.ac.uk/bsm/atlas>. Acessado em: 07/08/2012.

[7] Moléculas que contém grupos polares (hidrofílicos) e apolares (hidrofóbicos)

hidrofóbicas entre aminoácidos apolares podem estabilizar a estrutura tridimensional de uma proteína. São um dos maiores contribuintes para o enovelamento das proteínas, pois induzem a molécula a uma estrutura condensada reduzindo/evitando os contatos entre resíduos hidrofóbicos e moléculas de água (Lins e Brasseur, 1995; Pace *et al*, 1996).

Motta (2011) define interações hidrofóbicas como sendo forças não covalentes resultantes da tendência das cadeias laterais hidrofóbicas serem atraídas umas pelas outras, objetivando ocupar o menor volume possível, minimizando seus contatos com a água, que são liberadas do interior da molécula, aumentando a desordem do sistema. Lesser e Rose (1990) contabilizaram que 81% das cadeias laterais apolares (Ala, Val, Ile, Leu, Met, Phe, Trp, Cys), 70% dos grupos peptídicos, 63% das cadeias laterais polares (Asn, Gln, Ser, Thr, Tyr) e 54% das cadeias laterais carregadas (Arg, Lys, His, Asp, Glu) voltam-se para o interior da molécula durante o enovelamento, evitando contato com a água.

A Figura 11 representa uma fita dupla de DNA sendo enovelada e, sob forças hidrofóbicas, expulsando moléculas de água do seu interior.

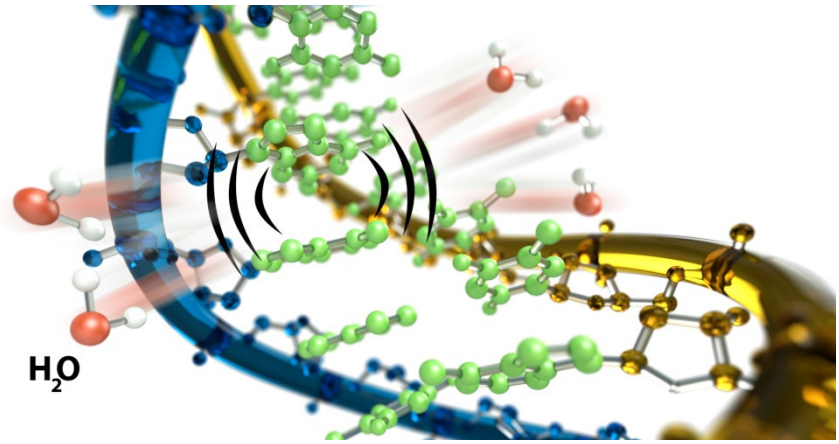
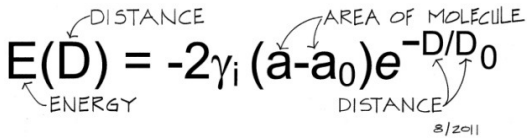


Figura 11 – Dupla hélice do DNA sob forças hidrofóbicas expulsando moléculas de água
Fonte: Werfhorst (2011), baseado no trabalho de Donaldson Jr. *et al*, 2011

Segundo Lins e Brasseur (1995), análises de hidrofobicidade de proteínas são realizadas através de análises estatísticas baseadas na hidrofobicidade de cada resíduo. A figura anterior foi extraída de uma matéria (Werfhorst, 2011) da Universidade Califórnia de Santa Bárbara abordando a publicação do Dr. Israelachvili (Donaldson Jr. *et al*, 2011), que propôs uma equação (Equação 4) para calcular a energia de uma interação hidrofóbica. Na equação, a energia hidrofóbica é proporcional à tensão interfacial (γ) e à área de superfície hidrofóbica

exposta (a – a_0), considerando também a distância (D e D_0) entre as moléculas. Pode ser aplicada, segundo os autores, até mesmo nos mais complicados sistemas como membranas celulares ou proteínas.



Equação 4 – Equação para o cálculo da energia de uma interação hidrofóbica

Fonte: Werfhorst (2011), publicada em Donaldson Jr. *et al* (2011)

1.2 Protein Data Bank

Existem inúmeras proteínas conhecidas e com sua estrutura tridimensional resolvida. Os bancos de dados biológicos buscam catalogá-las bem como armazenar o máximo de informações possíveis sobre sua estrutura, composição, função, dentre outras. Um exemplo de banco de dados público e gratuito é o *Protein Data Bank* - PDB (Berman *et al*, 2000), que detém, em sua atualização de 30/10/2012, 85.848 estruturas com um crescimento exponencial anual (Gráfico 1 - o gráfico original foi cortado para se exibir apenas os valores a partir do ano 2000).

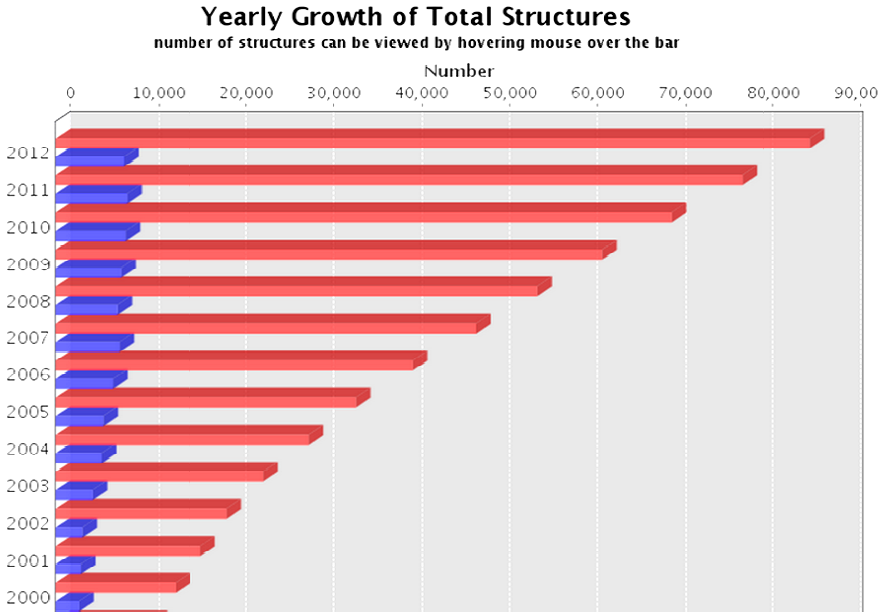


Gráfico 1 - Crescimento[8] anual do total de estruturas do PDB. Em azul o crescimento do ano, em vermelho o crescimento acumulado.

O PDB é gerenciado por dois membros do RCSB[9] (*Research Collaboratory for Structural Bioinformatics*): Rutgers (*The State University of New Jersey*) e UCSD (*University of California, San Diego*), e é financiado por NSF (*National Science Foundation*), NIGMS (*National Institute of General Medical Sciences*), DOE (*U.S. Department of Energy, Office of Science*), NLM (*U.S. National Library of Medicine, National Institutes of Health*), NCI (*National Cancer Institute*), NINDS (*National Institute of Neurological Disorders and Stroke*), e NIDDK (*National Institute of Diabetes and Digestive and Kidney Diseases*).

O PDB mantém, em sua base de dados, coordenadas atômicas e detalhes (ligações, distâncias, coeficientes, fatores, dentre outros) que envolvam a estrutura tridimensional de importantes macromoléculas biológicas como proteínas, DNA, RNA e complexos de proteínas e ácidos nucleicos. A Tabela 1 apresenta as quantidades mantidas de cada uma dessas estruturas considerando o método[10] experimental de resolução utilizado, podendo ser Cristalografia por Difração de Raios-X, Espectroscopia por Ressonância Magnética Nuclear (NMR) e Microscopia Eletrônica.

[8] PDB. Yearly Growth of Total Structures. Gráfico. - Disponível em: <http://www.pdb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100>. Acessado em: 03/11/2012.

[9] Site do grupo: http://home.rcsb.org

[10] Descrição dos métodos disponíveis no site do PDB em: http://www.pdb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/methods.html

Tabela 1 - Distribuição[11] das estruturas mantidas no PDB em sua atualização de 30/10/2012

Método experimental	Proteínas	Ácidos nucleicos	Complexos Proteína e Ácidos Nucleicos	Outros	Total
Raios-X	70547	1400	3556	3	75506
NMR	8455	1010	190	7	9662
Microscopia eletrônica	321	23	120	0	464
Híbrido	45	3	2	1	51
Outros	143	4	5	13	165
Total	79511	2440	3873	24	85848

1.2.1 Formato do arquivo PDB

As estruturas e suas informações são armazenadas no PDB através de *flat files*, ou seja, arquivos no formato texto, contendo toda a informação obtida a partir da resolução da estrutura, anotada pelo pesquisador que a resolveu. Esta estrutura, identificada por um código de 4 caracteres, é descrita neste arquivo em coordenadas atômicas, para cada um dos átomos que a compõe. O PDB disponibiliza 4 formatos diferentes de arquivo, sendo o formato arquivo PDB[12] (criado em 1970, usado e reconhecido por vários softwares, consistindo de dados provenientes de Difração de Raios-X ou estudos de NMR), mmCIF[13] (*macromolecular Crystallographic Information File*, descreve as entradas do PDB através de dicionários de dados), PDBML[14] (arquivo PDB em formato XML[15], envolvendo um *schema* XML para a definição e validação do arquivo) e *Chemical Component Dictionary*[16] (referência externa para descrever todos os resíduos e pequenas moléculas encontradas nas entradas do PDB).

Além da estrutura, há também detalhes da publicação gerada a partir da proteína, sua classificação, peso, identificação das cadeias, organismo, ligantes, domínio, detalhes sobre o método usado, resolução do arquivo gerado (em ângstrons), dentre outras informações. O formato de arquivo PDB é lido por várias ferramentas úteis ao pesquisador envolvido com proteínas como visualizadores de estrutura, softwares para sobreposição de cadeias, alinhamento, entre outros variados.

[11] Tabela disponível em http://www.pdb.org/pdb/statistics/holdings.do, acessada em 03/11/2012.

[12] Maiores informações em: http://www.wwpdb.org/docs.html

[13] Maiores detalhes em: http://mmcif.pdb.org/

[14] Maiores detalhes em: http://pdbml.pdb.org/

[15] eXtended Markup Language – Linguagem de marcação utilizada para identificar e transportar dados entre aplicações através de *tags* (marcadores) que delimitam o início e término da informação.

[16] Maiores detalhes em: http://www.wwpdb.org/ccd.html

A última versão 3.3[17] do formato padrão do arquivo PDB foi publicada em julho de 2011 com atualizações em outubro do mesmo ano e em maio de 2012. Segundo este documento, o arquivo PDB é apresentado como um conjunto de registros (6 primeiras colunas de cada linha), que identificam a informação que será disponibilizada naquela linha (cada linha tem no máximo 80 colunas). O arquivo é descrito em seções e cada seção tem seu conjunto de registros, conforme abordado na Tabela 2.

Tabela 2 - Registros do arquivo PDB[18] (tradução nossa)

Seção	**Descrição**	**Registros**
Identificação (*Title*)	Descrição da estrutura contida no arquivo	HEADER, OBSLTE, TITLE, SPLIT, CAVEAT, COMPND, SOURCE, KEYWDS, EXPDTA, NUMMDL, MDLTYP, AUTHOR, REVDAT, SPRSDE, JRNL
Anotações (*Remark Annotations*)	Comentários mais detalhados sobre a estrutura	REMARKs 0-999
Estrutura primária (*Primary structure*)	Sequência de peptídeos e/ou nucleotídeos	DBREF, SEQADV, SEQRES, MODRES
Heterogêneos (*Heterogen*)	Descrição de grupos fora do padrão	HET, HETNAM, HETSYN, FORMUL
Estrutura secundária (*Secondary Structure*)	Descrição da estrutura secundária contida no arquivo	HELIX, SHEET
Conectividade (*Connectivity annotation*)	Conectividade química	SSBOND, LINK, CISPEP, CONECT
Outras características (*Miscellaneous features*)	Características gerais da macromolécula	SITE
Cristalografia (*Crystallographic*)	Descrição da célula cristalográfica	CRYST1
Coordenadas de transformação (*Coordinate transformation*)	Operadores para a transformação de coordenadas	ORIGXn, SCALEn, MTRIXn
Coordenadas (*Coordinate*)	Coordenadas atômicas	MODEL, ATOM, ANISOU, TER, HETATM, ENDMDL
Finalização (*Bookkeeping*)	Informações de resumo ou finalização do arquivo	MASTER, END

Uma das seções mais importantes de um arquivo PDB é a seção de coordenadas atômicas, detalhada na Tabela 3, que apresenta as coordenadas e a identificação de cada átomo que compõe a estrutura registrada no arquivo. Estas coordenadas são usadas para a visualização tridimensional da estrutura bem como qualquer manipulação ou dinâmica que possa ser

[17] Disponível em: http://www.wwpdb.org/documentation/format33/v3.3.html

[18] Disponível em: http://www.wwpdb.org/documentation/format33/sect1.html

realizada com a proteína em questão (modelagem por homologia, sobreposição de estruturas, *docking*[19], dentre outros).

Tabela 3 – Formato da seção de coordenadas atômicas do arquivo PDB[20] (tradução nossa)

Colunas	Tipo de dado	Campo	Definição
1-6	Nome do registro	"ATOM "	
7-11	Inteiro	serial	Número serial do átomo
13-16	Átomo	name	Nome do átomo
17	Caracter	altLoc	Indicador de localização alternativa
18-20	Nome do resíduo	resName	Nome do resíduo
22	Caracter	chainID	Identificador da cadeia
23-26	Inteiro	resSeq	Número sequencial do resíduo
27	Caracter	iCode	Código para inserção de novos resíduos
31-38	Real (8,3)	x	Coordenada ortogonal para X em ângstrons
39-46	Real (8,3)	y	Coordenada ortogonal para Y em ângstrons
47-54	Real (8,3)	z	Coordenada ortogonal para Z em ângstrons
55-60	Real (6,2)	occupancy	Ocupância – probabilidade de o átomo estar naquela localização
61-66	Real (6,2)	tempFactor	Medida de confidencialidade da localização do átomo
77-78	String (2)	element	Símbolo do elemento, alinhado à direita
79-80	String (2)	charge	Carga do átomo

1.2.2 Exemplo de arquivo PDB

A Figura 12 apresenta um trecho do arquivo PDB 1IME[21] (conjuntos de linhas não exibidas foram substituídas por "..."). Na figura pode-se observar o grupo referente à estrutura (Transferase, registro HEADER) e sua identificação (registro TITLE), bem como de suas moléculas (registros COMPND e SOURCE, exibindo-se apenas uma molécula na figura). Detalhes sobre a forma de expressão da proteína (registros SOURCE 7 e 8) identificam ter

[19] Busca de um ligante candidato através da variação de sua conformação para aumentar o número de contatos.

[20] Disponível em: http://www.wwpdb.org/documentation/format33/sect9.html

[21] PDB ID 2IME. Thompson, L.C., Ladner, J.E., Codreanu, S.G., Harp, J., Gilliland, G.L., Armstrong, R.N. 2-Hydroxychromene-2-carboxylate Isomerase: a Kappa Class Glutathione-S-Transferase from Pseudomonas putida. Journal: (2007) *Biochemistry* 46: 6710-6722. PubMed: 17508726. DOI:10.1021/pdb2ime/pdb.

sido por um plasmídio (PET20B). Sua estrutura foi resolvida por Difração de Raios-X (registro EXPDTA). Outros detalhes que podem ser observados, dentre vários, é a publicação gerada a partir dessa estrutura (registros JRNL).

```
HEADER    TRANSFERASE                             04-OCT-06   2IME
TITLE     2-HYDROXYCHROMENE-2-CARBOXYLATE ISOMERASE: A KAPPA CLASS
TITLE    2 GLUTATHIONE-S-TRANSFERASE FROM PSEUDOMONAS PUTIDA
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: 2-HYDROXYCHROMENE-2-CARBOXYLATE ISOMERASE;
COMPND   3 CHAIN: A;
...
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: PSEUDOMONAS PUTIDA;
...
SOURCE   5 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
...
SOURCE   7 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE   8 EXPRESSION_SYSTEM_PLASMID: PET20B(+)
...
EXPDTA    X-RAY DIFFRACTION
...
JRNL        AUTH   L.C.THOMPSON,J.E.LADNER,S.G.CODREANU,J.HARP,
JRNL        AUTH 2 G.L.GILLILAND,R.N.ARMSTRONG
JRNL        TITL   2-HYDROXYCHROMENE-2-CARBOXYLIC ACID ISOMERASE: A
JRNL        TITL 2 KAPPA CLASS GLUTATHIONE TRANSFERASE FROM
JRNL        TITL 3 PSEUDOMONAS PUTIDA
JRNL        REF    BIOCHEMISTRY                 V.  46  6710 2007
JRNL        REFN                   ISSN 0006-2960
JRNL        PMID   17508726
JRNL        DOI    10.1021/BI700356U
...
```

Figura 12 - Trechos da identificação do arquivo PDB 2IME

Quanto à anotação do arquivo, pode-se observar na Figura 13 as resoluções alta (1,70 Å - Angstrons) e baixa (18,14 Å). O significado desses valores será dado na seção posterior (Resolução). Outros detalhes como o número de átomos de proteína (1702 átomos), ácidos nucleicos (0), heterogêneos (79) e átomos de solvente (161) podem ser obtidos bem como detalhes sobre o experimento que gerou o depósito como a data de realização (06/05/2004), a temperatura do ensaio (100º K), o pH usado na solução (6.1), o número de cristais (1) e o equipamento utilizado (Rigaku RU200).

```
...
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 1.70
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 18.14
...
REMARK   3  NUMBER OF NON-HYDROGEN ATOMS USED IN REFINEMENT.
REMARK   3   PROTEIN ATOMS            : 1702
REMARK   3   NUCLEIC ACID ATOMS       : 0
REMARK   3   HETEROGEN ATOMS          : 79
REMARK   3   SOLVENT ATOMS            : 161
...
REMARK 200 EXPERIMENTAL DETAILS
REMARK 200  EXPERIMENT TYPE                : X-RAY DIFFRACTION
REMARK 200  DATE OF DATA COLLECTION        : 06-MAY-04
REMARK 200  TEMPERATURE           (KELVIN) : 100
REMARK 200  PH                             : 6.1
REMARK 200  NUMBER OF CRYSTALS USED        : 1
...
REMARK 200  X-RAY GENERATOR MODEL          : RIGAKU RU200
...
```

Figura 13 - Trechos da anotação do arquivo PDB 2IME

A Figura 14 apresenta duas das 16 linhas que apresentam as sequências de resíduos da estrutura principal (registro SEQRES); o íon fosfato identificado como estrutura heterogênea de número 307, de 5 existentes, (registro HET), com sua identificação “PO4 Phosphate Ion” (registro HETNAM) e fórmula “PO4 3(O4 P 3-)” (registro FORMUL); 2 das 11 linhas que identificam as alfa hélices (registro HELIX); 2 das 4 que identificam as folhas beta (registro SHEET); uma única identificação do registro CISPEP, que informa a presença de uma valina (168) e uma prolina (169) encontradas em conformação cis, além dos registros CRYST1 (descrição da célula unitária), ORIGXn e SCALEn, que apresentam os operadores para a transformação de coordenadas.

```
...
SEQRES   1 A  203  MET ILE VAL ASP PHE TYR PHE ASP PHE LEU SER PRO PHE
SEQRES   2 A  203  SER TYR LEU ALA ASN GLN ARG LEU SER LYS LEU ALA GLN
...
HET    PO4  A 307       5
...
HETNAM     PO4 PHOSPHATE ION
...
FORMUL   2  PO4    3(O4 P 3-)
...
HELIX    1   1 SER A   11  GLY A   29  1                                  19
HELIX    2   2 ASP A   38  ILE A   46  1                                   9
...
SHEET    1   A 4 THR A  31  ALA A  36  0
SHEET    2   A 4 ILE A   2  PHE A   7  1  N  PHE A   5   O  ARG A  33
...
CISPEP   1 VAL A  168    PRO A  169          0        -5.80
...
CRYST1   71.126   75.833   38.301  90.00  90.00  90.00 P 21 21 2     4
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.014060  0.000000  0.000000        0.00000
SCALE2      0.000000  0.013187  0.000000        0.00000
SCALE3      0.000000  0.000000  0.026109        0.00000
...
```

Figura 14 – Trechos da estrutura primária, heterogêneos, estrutura secundária, conectividade, cristalografia e coordenadas de transformação do arquivo PDB 2IME

A seção de coordenadas atômicas do arquivo PDB, apresentada na Figura 15, permite observar os campos descritos na Tabela 3, como a identificação do átomo, do resíduo, bem como suas coordenadas atômicas. Interessante perceber na figura o primeiro resíduo da molécula (Metionina) e seus átomos da cadeia principal (N, CA, C, O) e os da cadeia lateral (CB, CG, SD, CE). A serina mostrada (resíduo 22) possui uma particularidade, seus carbono beta e oxigênio gama apresentam dupla conformação, ou seja, quando a estrutura foi resolvida foram encontradas duas posições para estes átomos que estão descritas no arquivo e identificadas com as letras C e D logo antes do nome do resíduo. A presença de um íon fosfato na molécula também pode ser observada através de suas coordenadas atômicas, pois os átomos de número 1709 a 1713 referem-se não a átomos de resíduos de aminoácidos mas aos átomos do íon fosfato (PO_4) presente na molécula.

Os vários métodos experimentais de resolução de estrutura possuem seus detalhes específicos, descritos nos campos citados acima, e neste documento foram ressaltadas algumas características, como resolução por exemplo, presentes apenas nos métodos de Difração por Raios-X, foco do trabalho desenvolvido.

```
...
ATOM      1  N   MET A   1      23.976  58.095  50.721  1.00 37.97           N
ATOM      2  CA  MET A   1      25.111  57.608  49.859  1.00 37.51           C
ATOM      3  C   MET A   1      24.516  57.054  48.565  1.00 34.95           C
ATOM      4  O   MET A   1      23.534  56.338  48.665  1.00 36.29           O
ATOM      5  CB  MET A   1      25.855  56.508  50.611  1.00 37.67           C
ATOM      6  CG  MET A   1      26.846  55.743  49.790  1.00 38.61           C
ATOM      7  SD  MET A   1      28.141  54.997  50.836  1.00 40.67           S
ATOM      8  CE  MET A   1      27.443  54.717  52.455  1.00 39.19           C
...
ATOM    189  N   SER A  22      20.286  45.145  46.471  1.00 23.12           N
ATOM    190  CA  SER A  22      18.944  45.438  46.950  1.00 25.26           C
ATOM    191  C   SER A  22      18.973  45.212  48.432  1.00 26.85           C
ATOM    192  O   SER A  22      18.326  45.952  49.170  1.00 26.66           O
ATOM    193  CB CSER A  22      17.955  44.477  46.308  0.50 24.47           C
ATOM    194  CB DSER A  22      17.890  44.551  46.309  0.50 24.76           C
ATOM    195  OG CSER A  22      18.089  43.187  46.872  0.50 23.89           O
ATOM    196  OG DSER A  22      16.604  44.979  46.752  0.50 25.44           O
...
HETATM 1709  P   PO4 A 307      -5.777  47.327  25.658  1.00 67.44           P
HETATM 1710  O1  PO4 A 307      -4.318  46.935  25.756  1.00 65.84           O
HETATM 1711  O2  PO4 A 307      -5.947  48.830  25.834  1.00 66.27           O
HETATM 1712  O3  PO4 A 307      -6.305  47.005  24.281  1.00 67.50           O
HETATM 1713  O4  PO4 A 307      -6.534  46.560  26.721  1.00 64.83           O
...
```

Figura 15 - Parte das coordenadas atômicas do arquivo PDB 2IME

1.2.3 Resolução

A resolução indicada em um arquivo PDB (valor indicado no arquivo como resolução alta – RESOLUTION RANGE HIGH -, gerado por experimentos de Difração de Raios-X), para Berman *et al* (2000), é a medida da qualidade dos dados coletados do cristal da proteína ou ácido nucléico, sendo a medida do nível de detalhamento presente no padrão de difração e o nível de detalhe que será medido no mapa de densidade eletrônica. Assim, resoluções altas (de até 1 Å, por exemplo) denotam a precisão da localização atômica no mapa de densidade eletrônica e, por outro lado, resoluções baixas (valores maiores que 3 Å) definem apenas o contorno da cadeia protéica, levando à inferência da estrutura atômica. A maioria das estruturas resolvidas cristalograficamente tem resolução entre estes dois valores. Os autores ainda afirmam que, como regra geral, há maior confiabilidade na localização de átomos em estruturas com valores de resolução menores, ou seja, estruturas de alta resolução. A Figura 16 apresenta a visualização da tirosina 103 de uma molécula de mioglobina, proveniente de dois arquivos PDB com resoluções diferentes. Observa-se que a imagem da esquerda, de maior resolução, apresenta maior detalhamento que a imagem da direita, de menor resolução. Este detalhamento pode ser observado pela variação da precisão da malha azul que contorna

regiões com alta densidade eletrônica. As duas imagens da figura foram geradas a partir de diferentes arquivos PDB, com as respectivas resoluções (1,0 Å e 2,7 Å) como descrito.

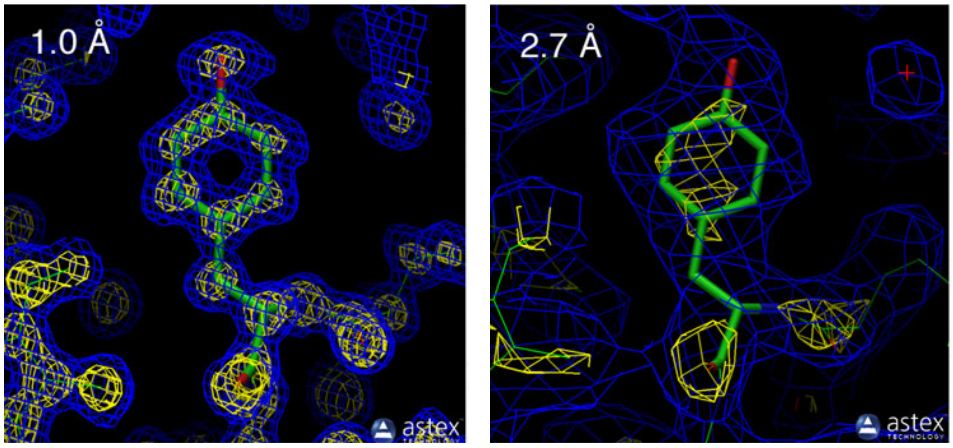


Figura 16 - Tirosina 103 da myoglobina, à 1Å (esquerda, PDB 1A6M[22]) e 2,7 Å (direita, PDB 108M[23])
Fonte: http://www.pdb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/resolution.html

[22] PDB ID 1A6M. Vojtechovsky, J., Chu, K., Berendzen, J., Sweet, R.M., Schlichting, I. Crystal structures of myoglobin-ligand complexes at near-atomicresolution. Journal: (1999) *Biophys Journal* 77: 2153-2174. PubMed: 10512835. PubMedCentral: PMC1300496. DOI: 10.1016/S0006-3495(99)77056-6. DOI:10.2210/pdb1a6m/pdb.

[23] PDB ID 108M. Smith, R.D. Correlations between Bound N-Alkyl Isocyanide Orientations and Pathways for Ligand Binding in Recombinant Myoglobins. Journal: (1999) Thesis, Rice. DOI:10.2210/pdb108m/pdb.

2. Justificativa

Hazes e Dijkstra (1988) desenvolveram o algoritmo SSBOND (SSBOND, 1999) com o intuito de identificar em uma proteína alvo (de estrutura tridimensional conhecida) pares de resíduos de aminoácidos que poderiam formar pontes dissulfeto caso estes mesmos resíduos fossem mutados para cisteínas. O algoritmo busca e classifica potenciais pares de resíduos de aminoácidos na proteína alvo observando a distância entre seus carbonos beta além dos ângulos diedrais. Assim, para cada resíduo na proteína alvo, posições de um provável Sγ (aminoácido mutado) são geradas matematicamente de forma que satisfaçam o requisito de que, com valores ideais para as distâncias Cα-Cβ e Cβ-Sγ e para o ângulo de ligação em Cβ, a distância entre o provável Sγ do resíduo 1 e o Cβ do resíduo 2 (dentro do par) seja igual ou muito próxima do valor ideal encontrado em pontes dissulfeto. Geralmente duas posições de Sγ são encontradas para cada cistina, resultando em até quatro diferentes conformações para uma ponte dissulfeto. Estas conformações são submetidas a procedimentos de minimização de energia e o usuário escolhe, ao final, aquela cuja conformação final seja a mais energeticamente aceitável.

Hazes e Dijkstra basearam-se nos trabalhos de Pantoliano *et al.* (1987) e de Wells e Powers (1986). Ambos estavam interessados no aumento da estabilidade autolítica da enzima *Subtilisina BPN'* com a introdução de pontes dissulfeto na estrutura terciária da proteína. Os últimos introduziram uma ponte dissulfeto entre os resíduos 24 e 87 da enzima, mas de acordo com os autores isto não afetou sua estabilidade em relação à autólise (Wells & Powers, 1986). Pantoliano *et al*, contudo, geraram outros trabalhos comprovando o aumento da estabilidade da mesma proteína através da introdução de pontes dissulfeto em outras regiões, bem como com outros tipos de mutação (Pantoliano *et al.*, 1988; 1989).

Almog *et al.* (2002) abordam as bases estruturais da termoestabilidade. A partir de duas variantes da enzima *Subtilisina BPN'* conseguiram um aumento de 1000 vezes na estabilidade

da proteína, fruto de 10 mutações pontuais, incluindo a introdução de uma nova ponte dissulfeto. Tigerström *et al.* (2004) descrevem ainda os efeitos da introdução de pontes dissulfeto e interações eletrostáticas via mutações sítio dirigidas na termoestabilidade da proteína *Azurina*.

Além de se verificar a importância da introdução de novas interações para a estabilidade de proteínas, pode-se optar pela remoção dessas interações e acompanhar o decaimento da estabilidade dessas moléculas. Sakaguchi *et al.* (2008) realizaram experimentos justamente com este enfoque. Ao mutarem dois resíduos de cisteína da proteína *Aqualysina I* (Cys99 e Cys194) para serinas, observaram não apenas o rompimento das pontes dissulfeto Cys67-Cys99 e Cys163-Cys194, como também uma perda significativa da termoestabilidade dessa proteína. Concluíram ainda que a ponte dissulfeto Cys163-Cys194 é mais importante para a atividade catalítica e estabilidade conformacional da *Aqualysina I* do que a Cys67-Cys99 (Sakaguchi *et al.*, 2008). Sakaguchi *et al.* (2007) também concluíram, para esta mesma proteína, que sua termoestabilidade estava relacionada aos resíduos de prolina nas regiões C- e N- terminal. Os parâmetros que definem as pontes dissulfeto estão armazenados em diversos arquivos no formato PDB (*Protein Data Bank*, Berman *et al*, 2000).

Kuroki, Weaver e Matthews (1993) estudaram a mutação da treonina 26 para ácido glutâmico na lisozima fase T4 (T4L) e seu efeito na parede celular da Escherichia coli, onde o resíduo de aminoácido inserido formou nova ligação covalente. Em trabalho posterior, Kuroki, Weaver e Matthews (1999) avaliaram a alteração da função da enzima na mutação da treonina 26 para histidina, também na lisozima fase T4 (T4L). Naquele trabalho, vários mutantes foram avaliados e uma tabela com 25 mutantes é apresentada contendo a variação da atividade de cada um deles a partir da mutação realizada. Em 2010, Matthews e outro grupo (Baase, Liu, Tronrud, Matthews, 2010) publicaram um trabalho extenso sobre a estrutura, estabilidade e enovelamento da lisozima fase T4. Nesse novo trabalho apresentaram uma tabulação completa de todos os variantes que foram caracterizados incluindo temperatura de *melting*, dados cristalográficos, códigos PDB, além das referências à literatura original. Um dos resultados encontrados pelo grupo é que a proteína é muito tolerante a mudanças na sequência de aminoácidos. Avaliaram também as situações em que ocorreu aumento da estabilidade da proteína a partir das mutações, em um dos casos a temperatura de melting foi aumentada em 23.4°C.

Sobre as ligações de hidrogênio, Krasil'nikov, Pashchenko e Noks (2001) concluíram que, como resultado da mutação sítio dirigida ou modificação da rede nativa de ligações de hidrogênio em alguns sítios ativos de um grupo de proteínas, houve uma perturbação na variação da energia ótima entre os estados inicial e final: o aumento do número de pontes de hidrogênio e a substituição isotópica $H_2O \rightarrow D_2O$ foi acompanhada pelo aumento da quantidade de energia gasta na solvatação. Pace *et al* (2001) também estudaram a contribuição da ligação de hidrogênio da tirosina na estabilidade protéica. Miyawaki e Tatsuno (2011) analisaram a importância do papel da perturbação das ligações de hidrogênio assim como das interações hidrofóbicas, como um mecanismo para a desestabilização de proteínas por alcoóis.

Deutsch e Krishnamoorthy (2007) desenvolveram uma função para predizer os efeitos de mutações simples ou múltiplas na estabilidade e reatividade de proteínas. Esta função, além dos inúmeros resultados baseados em mutações variadas já estudadas na Literatura, indica a necessidade de um mecanismo eficiente para identificar possíveis mutações "aceitáveis" em uma proteína alvo. Entende-se por aceitável aquele mutante "*in silico*" que tem possibilidade estereoquímica de existir "*in vitro*". Até onde sabemos, não existe uma base de dados das conformações das cadeias principais dos pares de resíduos de aminoácidos interagentes (que interagem por interação iônica, ponte de hidrogênio ou ponte dissulfeto) que possa ser utilizada para a proposição de duas mutações concomitantes em proteínas alvo baseando-se na conformação de sua cadeia principal.

Uma base de dados como esta pode ser usada para a proposição de mutações em proteínas de estrutura tridimensional conhecida de forma a propiciar a formação de novas interações visando um aumento da termoestabilidade e/ou estabilidade conformacional de uma determinada proteína. Além disto, o fato de propor mutações em resíduos de aminoácidos que mantêm conformações das cadeias principais similares àquelas encontradas no banco de dados, sugere que o enovelamento da proteína alvo pode ser mantido e possivelmente a sua função. É claro que a mudança das cadeias laterais pode influenciar drasticamente na conformação da proteína alvo, como já se sabe da experiência (e.g. hemoglobinas tipo S). Contudo, garantir a conformação das cadeias principais, é diminuir os riscos de uma mutação gerar a mudança de conformação na proteína alvo.

Vários algoritmos e softwares têm sido desenvolvidos para casos específicos de mutações ou de predição de interações como, dentre outros:

- SSBOND[24] (1999), um servidor para predição de pontes dissulfeto. Dado um arquivo PDB, retorna uma lista de pares de resíduos que, se mutados para cisteínas, formarão pontes dissulfeto.

- DiANNA[25] - *DiAminoacid Neural Network Application* - (Ferrè, Clote; 2005b), servidor que, a partir de uma sequência no formato FASTA, determina o estado de oxidação da cisteína e a conectividade para ponte dissulfeto de uma proteína, a partir de sua sequência de aminoácidos.

- DISULFIND[26] (Ceroni *et al*; 2006), prediz o estado de ligação das cisteínas e sua conectividade a partir de uma sequência no formato FASTA. Prevê padrões de ponte dissulfeto para dois estágios computacionais: 1) ponte dissulfeto prevista pelo classificador *BRNN-SVM binary classifier*; 2) cisteínas conhecidas por participarem de ponte, reconhecidas por rede neural.

- DIpro[27] (Cheng, Saigo, Baldi; 2006), baseado em rede neural, support vector machine, graph matching e algoritmos de regressão. Funciona a partir de uma sequência no formato FASTA e retorna o número de pontes e o estado de conectividade de cada cisteína.

- EDBCP[28] - *Ensemble-based Disulfide Bonding Connectivity Pattern prediction server* - (Lin, Tseng; 2010), também faz a previsão de pontes dissulfeto a partir de sequência no formato FASTA.

- HBOND[29] (Mizuguchi *et al*, 1998), localiza possíveis ligações de hidrogênio na estrutura da proteína. Usa como entrada o arquivo no formato PDB, que tem que

[24] Disponível em: <http://eagle.mmid.med.ualberta.ca/forms/ssbond.html>, porém não mais acessível em 10/06/2010.

[25] Disponível em: <http://clavius.bc.edu/~clotelab/DiANNA>, acessado em 10/05/2010.

[26] Disponível em: < http://cassandra.dsi.unifi.it/disulfind>, acessado em 05/07/2010

[27] Disponível em: <http://contact.ics.uci.edu/bridge.html>, acessado em 10/05/2010, porém não funcional, a submissão enviada não é respondida por e-mail, conforme se propõe a ferramenta.

[28] Disponível em: <http://120.107.8.16/dbcp>, acessado em 20/06/2010.

atender a uma série de requisitos para a execução da ferrramenta como não ter dupla conformação, átomos de hidrogênio, dentre outros.

- COILCHECK[30] (Alva, Syamala, Sowdhamini, 2008), servidor para cálculo de várias interações dentro do arquivo PDB submetido pelo formulário online. Calcula ligações de hidrogênio, interações hidrofóbicas, pares de van der Waals, pontes salinas, além da energia das ligações de hidrogênio, eletrostáticas, de van der Waals. A ferramenta apenas identifica as interações e calcula as distâncias e as energias entre os átomos envolvidos.

Além desses, outros trabalhos também são relevantes como de Sowdhamini *et al* (1989), com a definição de vários critérios para caracterizar pontes dissulfeto; Muskal *et al* (1990), com a predição dos estados de ligação das cisteínas nas cadeias; Bhattacharyya, Pal e Chakrabarti (2004), com a análise do ambiente estereoespecífico e a conservação das pontes nas estruturas protéicas; Tsai *et al* (2005), que, através de *support vector machines* (SVM), propõem a predição de pontes utilizando o valor da distância sequencial entre cisteínas oxidadas, Singh (2008), que faz uma revisão das técnicas algorítmicas para determinação de pontes dissulfeto. Em trabalho anterior Singh e Thornton (1992) elaboraram um atlas das interações das cadeias laterais das proteínas. O atlas é composto de 2 volumes e 400 possíveis pares de interação entre os 20 tipos diferentes de aminoácidos, que são apresentados em forma de tabelas com dados estatísticos, histogramas e representações tridimensionais dos clusteres gerados a partir de 62 estruturas de alta resolução depositadas no PDB. Um atlas similar a este foi desenvolvido por McDonald e Thornton (1994) abordando apenas as ligações de hidrogênio, tanto da cadeia principal quanto da cadeia lateral.

Apesar de todos estes trabalhos aqui citados, não há na Literatura uma base de dados de pares de aminoácidos interagentes que possibilite a introdução sistemática de mutações que propiciarão a formação de novas interações (ponte dissulfeto, interação iônica e ponte de hidrogênio) entre os aminoácidos mutados.

---

[29] Disponível em: <http://caps.ncbs.res.in/iws/hbond.html >, acessado em 10/01/2011.

[30] Disponível em: <http://caps.ncbs.res.in/coilcheckplus/>, acessado em 10/01/2011.

A estabilidade adquirida com essas mutações possibilitaria a uma proteína atuar em ambientes mais agressivos como na indústria petroquímica, alimentícia, dentre outras, onde existem grandes variações no meio, como temperatura e pH, por exemplo.

3. Objetivos

3.1 Objetivo geral

Busca de pares de resíduos de aminoácidos interagentes, em proteínas de estrutura tridimensional conhecida, e identificação de padrões das conformações das cadeias principais dentro desses pares, com objetivo de gerar um banco de dados para propor duas mutações concomitantes em uma proteína alvo que possibilitem a ela adquirir maior estabilidade térmica ou química, via formação de novas interações (ligações de hidrogênio, pontes dissulfeto, interações iônicas); mas que não impliquem em alterações funcionais e de enovelamento.

3.2 Objetivos específicos

I. Buscar e analisar vários critérios para a definição de interação entre resíduos de aminoácidos (ponte dissulfeto, ligação iônica e ponte de hidrogênio) propostas na Literatura a fim de se estabelecer parâmetros mínimos e máximos que caracterizem uma interação. Por hora, trataremos apenas das interações entre átomos de cadeias laterais distintas ou entre um átomo de uma cadeia lateral com outro de uma cadeia principal de outro resíduo[31].

[31] As interações entre os átomos de cadeias principais distintas não serão analisadas, uma vez que por definição do projeto, tais interações não seriam incluídas ou retiradas nos mutantes gerados.

II. Varrer várias estruturas bem definidas (de alta resolução) de proteínas conhecidas identificando as interações entre pares de resíduos seguindo os critérios definidos no item I.

III. Montar uma base de dados (estrutural) com os pares de resíduos interagentes encontrados nas várias proteínas pesquisadas gerando arquivos no formato PDB individuais para cada interação.

IV. Realizar a curagem da base de dados a fim de excluir as interações que sejam idênticas considerando a sobreposição dos átomos no espaço.

V. Classificar e clusterizar as interações resultantes de forma a direcionar as buscas na base de dados. Estes processos podem ser feitos tanto do ponto de vista estrutural (sobreposição de cadeias principais) quanto bioquímico (tipos de resíduos envolvidos, tipos de átomos envolvidos).

VI. Avaliar o banco de dados acurado e identificar valores máximos e mínimos de distâncias entre átomos da cadeia principal dos resíduos interagentes.

VII. Selecionar em uma proteína alvo, pares de resíduos que satisfaçam os critérios definidos no item VI.

VIII. Comparar cada par encontrado no item VII contra o banco de dados (em relação à cadeia principal dos resíduos envolvidos) e identificar possíveis mutações que implicarão na formação de uma nova interação entre os resíduos mutados.

IX. Gerar um sistema a ser disponibilizado via Internet, para a manipulação da base de dados desenvolvida, bem como a busca de possíveis mutantes em uma proteína alvo, seguindo a metodologia aqui descrita.

4. Metodologia

O projeto foi desenvolvido somente com softwares livres, sendo utilizados a linguagem de programação PHP[32] (versão 5.3.2), o interpretador de comandos *Bash Shell*, um banco de dados em Mysql[33] (versão 5.1.47), servidor web Apache[34] (versão 2.2.14-5ubuntu8), utiliza ainda o aplicativo lsqkab do pacote CCP4[35] (versão 6.1.3), sendo executado num servidor (Core 2 Duo 2,8 GHz – 2 cpus, 4 GB Ram, 1,5 TB HD Raid 1) com sistema operacional Linux Fedora[36] Core 12 (kernel 2.6.32.9-70.fc12.i686.PAE). Parte foi executada no cluster de computadores Veredas, do CENAPAD (Centro Nacional de Processamento de Alto Desempenho) da UFMG, composto por 106 nós de processamento idênticos, cada nó com dois processadores quad-core Intel Xeon X5355 e 16,0 Gigabytes de memória principal, totalizando 848 cores e aproximadamente 1,7 Terabytes de memória distribuída.

Todos os procedimentos para geração dos arquivos das interações e dos candidatos da proteína alvo, bem como inserção desses dados no banco de dados, foram implementados através de Shellscripts, que consistem em arquivos executáveis contendo comandos e chamadas de programas executáveis via prompt de comando. A implementação usando *shellscripts* (*bash shell*) permitiu o uso de funções e programas específicos de manipulação de arquivos e processos do sistema operacional que otimizaram a geração dos arquivos das interações, que compõem o banco.

O uso dos *shellscripts* também proporcionou melhor controle da execução da aplicação de sobreposição de estruturas tridimensionais, lsqkab, e o uso de execução paralela de processos. Esta última consistiu na execução simultânea de vários processos para geração dos arquivos

[32] Disponível em http://www.php.net
[33] Disponível em http://www.mysql.com
[34] Disponível em http://www.apache.org
[35] Disponível em http://www.ccp4.ac.uk
[36] Disponível em http://www.fedoraproject.org

das interações ou das sobreposições, reduzindo também o tempo de execução dessa tarefa de uma forma geral.

Segundo Neves (2006), Shellscript não é uma linguagem de programação, mas um arquivo executável contendo comandos do *Shell* (interpretador) do sistema operacional. Jargas (2008) acrescenta que apesar de não ser uma linguagem possui comandos de iteração e condicionais como *for*, *while*, *if*, além de variáveis e funções, para permitir a execução ou manipulação de arquivos e executáveis, necessárias na execução de *scripts*, como por exemplo a execução de um comando em segundo plano, com o uso do "&" ao final da linha de comando ou mesmo uma execução sem a necessidade da manutenção do shell ativo com o comando nohup. Para um comando ou script que se manterá executando por muito tempo e não havendo a necessidade do usuário permanecer conectado na máquina.

Os principais comandos, funções e programas usados nos *shellscripts* foram:

- egrep – uso de expressões regulares para a busca de padrões em arquivo texto, no caso, na Figura 17 linha 1 é listada apenas a primeira linha do arquivo contendo os parâmetros descritos na expressão regular que segue o comando.

- awk – muito útil na extração de informações dos nomes dos arquivos, com o uso principalmente de substr, print, split, como no exemplo a seguir (Figura 17, linha 2) que monta o nome do diretório (dir) a partir da remoção do prefixo "tmp-" do nome do diretório ($dirbonds). Na mesma figura observa-se o uso do pipe "|" – canalizador de saída permitindo que a saída de um comando se torne entrada do seguinte, sem a necessidade de geração de arquivos intermediários para isso.

- cat – leitura de arquivo em memória, permitindo a manipulação dos seus dados (Figura 17, linha 2).

- bc – juntamente com o uso de *scale* e *echo* (impressão na saída padrão, Figura 17, linhas 3, 5 e 7) permite o cálculo matemático com casas decimais, o que não é permitido pelas operações básicas de *Shell*. O parâmetro passado com o *scale* permite definir quantas casas decimais serão usadas. (${ref[$y]}) é a forma de acessar um vetor ($ref) na posição $y.

- concatenação de strings e resultados, como na linha 7, que gerará uma linha contendo dois números separados por espaço seguidos do nome de um arquivo.

- sed – editor de stream (fluxo de texto) usado para a substituição de espaço seguido de ponto " ." por zero seguido de ponto "0.", no trecho `sed 's/ \./0./g'` (Figura 17, linha 8). Juntamente com o sed também foi usado o tr que permite a substituição de um caracter, na linha, ":" por "\n".

- sort – para ordenação de resultados, no caso da linha 8, ordenação numérica (parâmetro g). Outro item interessante usado é o redirecionamento de saída ">" que permite gravar em um arquivo o resultado gerado pela sequência de comandos executada na linha.

```
1   line1=`egrep "^ATOM +[0-9]+ +$atom1 *[A]?$residue1 *$chain *$residue1num +\-
    ?[0-9]+\.[0-9]+ " $pdbdir/$file | head -n1`
2   matrix=`cat $dir/deltas/$file | awk '{ print $1 }' | tr '\n' ' '`
3   tmp=`echo "scale=8;($x-(${ref[$y]}))^2+$tmp"|bc`
4   for x in $matrix; do
5         tmp=`echo "scale=8;($x-(${ref[$y]}))^2+$tmp"|bc`; (( y++ ))
6   done
7   dist[$i]=`echo "scale=8; sqrt($tmp)"|bc`" $i $file:"; (( i++ ))
8   echo ${dist[@]} | tr ':' '\n' | sed 's/ \./0./g' | sed '/^$/d' | sort -g >
    $dir/dist.array
```

Figura 17 - Trecho de código *shellscript* usado no *script* de geração dos arquivos das interações

Uma pesquisa bibliográfica extensa, com seus resultados apresentados na Tabela 5, Tabela 6 e Tabela 7, foi utilizada para obtenção de parâmetros geométricos (distâncias) e químicos (tipo de átomo e carga) que definem as classificações das interações entre os átomos de um par de resíduos de aminoácidos (eletrostática – seção 1.1.3.1, ponte dissulfeto – seção 1.1.3.2, ligação de hidrogênio – seção 1.1.3.3, hidrofóbica – seção 1.1.3.4), de forma a se definir os critérios que foram usados durante a busca dessas interações nas estruturas das proteínas depositadas no *Protein Data Bank*. Estes critérios envolvem a direção da busca na cadeia, as distâncias interatômicas a serem consideradas, o método experimental usado bem como a resolução obtida.

A Figura 18 demonstra o fluxo de funcionamento do sistema, abordando a interação entre usuário, adminstrador, banco de dados, interface web e repositório de arquivos (HD). O administrador, através de shellscripts gera os arquivos com os das interações atômicas a partir

dos arquivos do PDB, os salva no disco e insere estas interações no banco de dados MySQL, que será acessado através da interface web pelo administrador ou usuário.

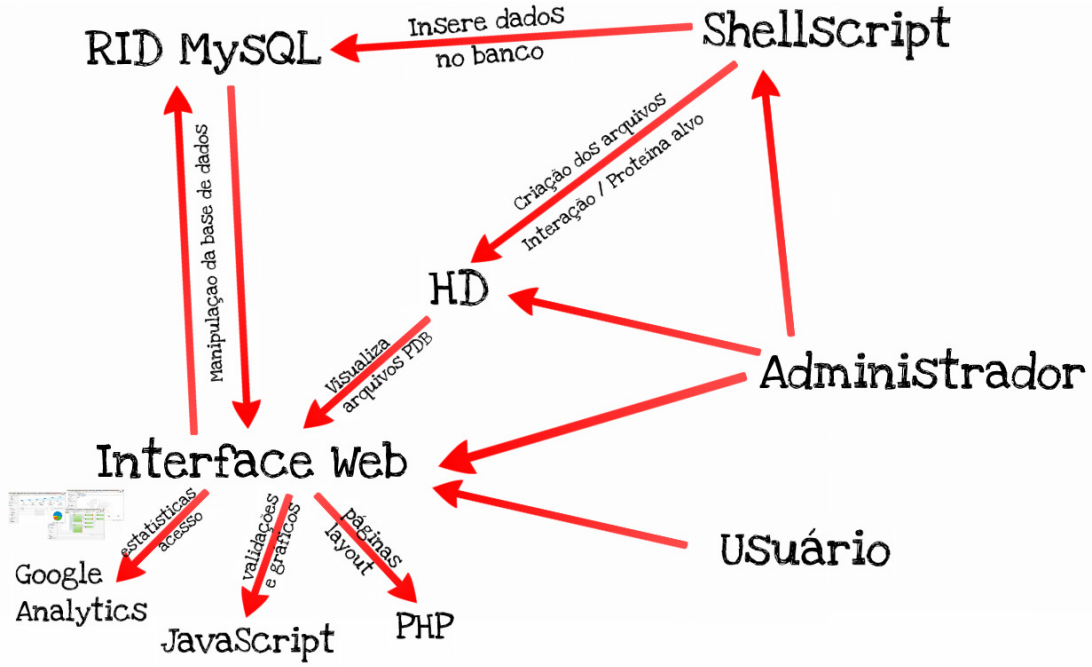


Figura 18 - Diagrama do funcionamento do sistema

4.1 Banco de dados

O banco de dados PDB teve sua estrutura explorada com o objetivo de se definir as informações e a forma de busca das mesmas dentro da estrutura do banco, ou seja, nos *flat files*. A estrutura do arquivo PDB foi usada para a geração de novos arquivos baseados no arquivo original da proteína, de forma a se isolar a interação encontrada, para a montagem de nova base de dados.

Após a definição desses parâmetros e estruturas de arquivos, uma base de dados estrutural foi montada com os pares de resíduos de aminoácidos encontrados (que satisfazem estes critérios) durante varredura do número máximo possível de proteínas do PDB. Estas proteínas foram selecionadas dentre as que atentassem a critérios como método (Difração por Raios-X) e resolução (valor da resolução alta <= 2 Å), o Algoritmo 1 apresenta a lógica que foi usada para a montagem do banco de interações.

Algoritmo 1 – Montagem do banco de interações

```
Download dos arquivos PDB resolvidos por Difração de Raios-X com resolução <= 2
Para cada interação a ser inserida no banco faça
    residue1 ← Três letras que identificam o resíduo 1
    atom1 ← Três letras que identificam o átomo 1
    residue2 ← Três letras que identificam o resíduo 2
    atom2 ← Três letras que identificam o átomo 2
    cutoff ← Valor da distância a ser observada entre os átomos
    type ← Tipo da interação
    // (HB – Hydrogen Bond, DB – Disulfide Bridge, IE – Interação Eletrostática)
    Para cada arquivo PDB faça
         Para cada cadeia do arquivo faça
              Para cada residuet1 = residue1
                   Para cada residuet2 = residue2 e residuet2num > residuet1num
                     Se distância (residuet1, residuet2) <= cutoff
                     Então GeraArquivoInteracao(residuet1, residuet2)
                          InsereInteracaoNoBanco()
                     FimSe
                   FimPara
              FimPara
         FimPara
    FimPara
FimPara
```

Uma vez criada a base de interações, uma busca de padrões entre os átomos que compõem as cadeias principais dos aminoácidos interagentes (do banco de dados) foi feita. A definição de propriedades geométricas (distâncias) entre os átomos foi utilizada como critério para busca de pares de aminoácidos próximos em proteínas-alvo, permitindo indicá-los como possíveis candidatos à mutação.

Algoritmo 2 – Levantamento das distâncias

```
Para cada arquivo de interação
    Para cada Atom1 em (N, CA, C, O)
         Para cada Atom2 em (N, CA, C, O)
              distAtom1Atom2 ← distância entre os átomos Atom1 e Atom2
              InsereDistanciaAtom1Atom2InteracaoNoBanco
         FimPara
    FimPara
FimPara
```

A função de inserção de distância no banco grava numa tabela todas as 16 distâncias possíveis entre os átomos (N, Cα, C, O) dos dois resíduos interagentes associadas a determinada

interação. A distância é calculada através do método Euclidiano, ou seja, a distância é igual à raiz quadrada da soma dos quadrados das diferenças entre as coordenadas atômicas (Equação 5).

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

Equação 5 – Fórmula utilizada para o cálculo da distância (euclidiana)

4.2 Busca

Uma busca direta na base de dados de interações consiste na sobreposição do par candidato, proveniente da proteína alvo, com todos os pares da interação até se encontrar um par da interação que tenha a melhor sobreposição atômica (menor distância entre os átomos pares, tendendo a zero implica em sobreposição perfeita). Os pares candidatos da proteína alvo são obtidos a partir do Algoritmo 3, consistindo de uma pesquisa de qualquer par de resíduos que atendam às distâncias obtidas no Algoritmo 2.

Algoritmo 3 – Busca de candidatos na proteína alvo

```
Download do arquivo PDB da proteína alvo
Para cada interação a ser pesquisada na proteína faça
    Para cada cadeia do arquivo faça
        Para cada residuet1 da cadeia
            Para cada residuet2 da cadeia e residuet2num > residuet1num
                Para cada Atom1t em (N, CA, C, O) de residuet1
                    Para cada Atom2t em (N, CA, C, O) de residuet2
                        distAtom1tAtom2t ← distância entre os átomos Atom1t e Atom2t
                        distAtom1Atom2 ← distância entre os átomos Atom1 e Atom2
                        Se distAtom1tAtom2t <= disAtom1Atom2 da interação
                        Então GeraArquivoParCandidato(residuet1, residuet2)
                            InsereParCandidatoNoBanco()
                        FimSe
                    FimPara
                FimPara
            FimPara
        FimPara
    FimPara
FimPara
```

Dada a forma de disponibilização dos dados nesta base de dados (*flat files*) houve a necessidade de se definir uma forma de busca que fosse mais eficiente que a sobreposição dos pares (Algoritmo 4), uma vez que uma comparação na base implicaria numa sobreposição atômica das estruturas a serem comparadas. Para a busca de um par da proteína-alvo no banco de interações se torna necessária a comparação dessa estrutura com cada uma das presentes no banco, através de sobreposição atômica, gerando como resultado um arquivo (arquivo de deltas) contendo as distâncias entre dois átomos pares pertencentes aos dois resíduos (N-N, Cα-Cα, C-C, O-O). Esta sobreposição, por ser tarefa que exige muito processamento e operações de escrita e leitura de arquivo em disco, torna a busca inviável, demandando então uma nova forma de busca definida neste trabalho.

Algoritmo 4 – Busca básica de interações numa proteína

```
Para cada interação a ser pesquisada na proteína faça
    Para cada par candidato encontrado nesta interação faça
        Para cada par da interação faça
            ExecutaSobreposicao (par proteína alvo, par interação)
            Ordena arquivos internamente da maior para a menor distância
            Ordena lista de arquivos com a maior distância encontrada
            Primeiros pares da lista são fortes candidatos à mutação
        FimPara
    FimPara
FimPara
```

A otimização da busca consistiu no cálculo de um *score* a partir dos arquivos de deltas (Equação 6), com o objetivo de ser usado como identificador/assinatura, para buscas mais rápidas que as realizadas através de sobreposições individualizadas de estruturas. Este cálculo é baseado na distância euclidiana, método muito comum na busca e avaliação de similaridades, e consiste na raiz quadrada das somas dos quadrados de cada par de diferença entre os respectivos valores de delta dos dois arquivos comparados.

$$d = \sqrt{\sum_{i=1}^{n} \left| x_i - y_i \right|^2}$$

Equação 6 – Equação para cálculo do valor do *score*, baseado na distância euclidiana

Assim, dado um conjunto de arquivos de pares da interação gerados, é feita uma sobreposição de cada um deles com o de maior resolução (Algoritmo 5), calculando-se o *score* definido anteriormente. O mesmo se aplica aos arquivos dos pares candidatos da proteína alvo.

Algoritmo 5 – Otimização dos pares da interação

```
Para cada interação faça
    referência_interação ← par interação maior resolução
    Para cada par da interação faça
        ExecutaSobreposicao (par interação, referência_interação)
        Ordena arquivo internamente da maior para a menor distância
    FimPara
FimPara
Ordena lista de todos os arquivos identificados pela maior distância encontrada
```

A partir das otimizações geradas, obtém-se dois vetores de valores calculados da mesma forma que podem refletir consequentemente a mesma situação. Ou seja, dado um valor no vetor de pares de proteínas, o mesmo valor no vetor de pares da interação pode levar a arquivos idênticos, com sobreposição perfeita, um possível mutante. Por outro lado, o cálculo pode gerar um valor igual porém a realidade do arquivo pode ser bem diferente. A busca portanto se dará através de uma pesquisa neste vetor de valores a partir do valor calculado para o par candidato da proteína alvo (Algoritmo 6).

Algoritmo 6 – Busca de interação na proteína

```
Para cada interação faça
    Para cada par candidato da proteína faça
        ExecutaSobreposicao (par proteína, referência_interação)
        Ordena arquivo internamente da maior para a menor distância
    FimPara
    Busca valor igual ou próximo ao obtido no vetor de valores dos pares de interação
FimPara
```

Assim, dada uma estrutura bem definida de uma proteína-alvo, para cada par de resíduos de aminoácidos dessa proteína-alvo (que satisfaçam os critérios geométricos estabelecidos anteriormente), será feita uma busca na base de dados a fim de se obter uma conformação das cadeias principais dos resíduos do par que sejam similares (provavelmente apresentando melhores sobreposições espaciais) à conformação das cadeias principais dos resíduos da

proteína-alvo. Para cada par encontrado na base, duas possíveis mutações podem ser propostas na proteína-alvo com o objetivo de se introduzir esta nova interação na estrutura analisada.

Além do banco de dados, para armazenamento dos pares de resíduos interagentes, há a necessidade de um sistema para interação humana nesta busca. Assim, um sistema com alta disponibilidade e acesso online, sem a necessidade de instalação de software adicional se torna interessante neste momento. Este sistema foi desenvolvido para permitir o uso da busca por mutantes a partir de outros pesquisadores.

5. Resultados

A implementação do código foi realizada em módulos de forma que possam ser alterados sem que os resultados anteriores sejam prejudicados. Cada etapa gera uma nova tabela no banco de dados e/ou arquivos no formato PDB com trechos do arquivo original ou no formato texto comum. As etapas serão descritas a seguir juntamente com as análises dos seus resultados parciais.

O trabalho aqui apresentado foi dividido em duas etapas, a saber: desenvolvimento da base de dados e desenvolvimento do sistema. A primeira etapa refere-se à base de interações que será utilizada para a proposição de mutações. Já a segunda etapa, refere-se ao desenvolvimento de interface web para possibilitar a busca de mutações possíveis a partir de uma proteína alvo.

5.1 Desenvolvimento da base de dados

Para a inserção de interações foi necessário caracterizar a interação na Literatura (Tabela 6, Tabela 7), para identificar os átomos que as realizam, como indicado por Mancini et al (2003). A Tabela 4 apresenta esta sequência de átomos e resíduos, na ordem em que devem ser pesquisados nos arquivos das proteínas, considerando apenas as interações que ocorrem entre cadeias laterais, conforme o objetivo desse trabalho. Além disso, apresenta também a quantidade de interações inseridas para cada tipo e o *cutoff* (distância entre os átomos) utilizado.

Tabela 4 - Interações e suas características

	Tipo de Interação	Grupos de Resíduos e Átomos Interagentes		*Cutoff*	Quantidade de interações inseridas
1	**Ponte Dissulfeto**	CYS SG	CYS SG	2,4 Å	1
2	**Ligação de Hidrogênio** (lateral positivo - lateral quase doador)	LYS NZ HIS ND1 HIS NE2 ARG NH1 ARG NH2	SER OG THR OG1 TYR OH	3,2 Å	30
3	**Ligação de Hidrogênio** (lateral positivo - receptor lateral)	LYS NZ HIS ND1 HIS NE2 ARG NH1 ARG NH2	ASN OD1 GLN OE1	3,2 Å	20
4	**Ligação de Hidrogênio** (doador lateral – lateral quase doador)	TRP NE1 ASN ND2 GLN NE2	SER OG THR OG1 TYR OH	3,2 Å	18
5	**Ligação de Hidrogênio** (doador lateral – receptor lateral)	TRP NE1 ASN ND2 GLN NE2	ASN OD1 GLN OE1	3,2 Å	12
6	**Ligação de Hidrogênio** (doador lateral – lateral negativo)	TRP NE1 ASN ND2 GLN NE2	ASP OD1 ASP OD2 GLU OE1 GLU OE2	3,2 Å	24
7	**Ligação de Hidrogênio** (lateral quase doador – receptor lateral)	SER OG THR OG1 TYR OH	ASN OD1 GLN OE1	3,2 Å	12
8	**Ligação de Hidrogênio** (lateral quase doador - lateral negativo)	SER OG THR OG1 TYR OH	ASP OD1 ASP OD2 GLU OE1 GLU OE2	3,2 Å	24
9	**Interação Eletrostática Atrativa** (lateral negativo - lateral positivo)	ASP OD1 ASP OD2 GLU OE1 GLU OE2	LYS NZ HIS NE1 HIS NE2 ARG NH1 ARG NH2	6,0 Å	40
			Total de interações:		181

No primeiro estágio do projeto foram consideradas somente as interações por ponte dissulfeto, para geração de uma base de dados inicial e testes visando à confirmação do método proposto. Estas interações foram caracterizadas pela distância máxima de 2,4 Angstrons (Å) entre os átomos de enxofre (enxofre gama) dos resíduos de cisteína. Esta medida foi definida a partir da distribuição das distâncias das pontes dissulfeto anotadas nos arquivos PDB demonstrada no Gráfico 4 e no Gráfico 5, além das várias medidas encontradas na Literatura para caracterizar as interações, conforme pode ser visto na Tabela 5.

Tabela 5 - Caracterização da distância da ponte dissulfeto

Distância	Publicação	Observações
3 Å	Martin *et al* (2011)	
2,15 Å	Singh (2008)	
2,04 Å	Ferrè e Clote (2005a) e Ferrè e Clote (2005b)	Com desvio padrão de 0,105 e distância máxima de 2,93 Å
2,3 Å	Bhattacharyya, Pal, Chakrabarti (2004)	descartando os maiores valores encontrados nos arquivos PDB variando de 2,31 Å a 2,89 Å
2,03 Å	Boisbouvier et al (2000)	
2,04 Å	Morris *et al* (1992)	com desvio padrão de 0,16
2,04 Å	Sowdhamini *et al* (1989)	também definem as distâncias entre os carbonos alfa (Cα-Cα, ≤ 6,5 Å), carbonos beta (Cβ-Cβ, ≤ 4,5 Å), entre o carbono beta e o enxofre (Cβ-S, ≤ 1,87 Å), bem como os ângulos diedros envolvidos na ponte

A segunda interação que fez parte da base foi a ligação de hidrogênio cujas características, observadas na Literatura, estão na Tabela 6. Optou-se por utilizar como distância de *cutoff* para esta interação a distância de 3,2 Å entre doador e aceptor de próton, pois como caracterizada por Jeffrey (1997), até esta distância encontram-se as ligações fortes e moderadas, estando também dentro da faixa da maioria dos autores apresentados na Tabela 6. Os doadores e aceptores de hidrogênio são os átomos O e N, conforme Figura 9, exibida anteriormente.

Tabela 6- Caracterização da distância da ligação de hidrogênio. As distâncias são entre doador e aceptor.

Distância	Publicação	Observações
3,6 Å	Guerois, Nielsen, Serrano (2002)	
2,4 – 3,5 Å	Eswar e Ramakrishnan (2000)	
2,2 – 2,5 Å forte **2,5 – 3,2 Å moderada** **3,2 – 4,0 Å fraca**	Jeffrey (1997)	Apesar de afirmar que uma distância média seria de 3 Å, ressalta que pode haver uma ligação de hidrogênio significante a uma distância de 3.5 Å entre doador e aceptor.
2,4 Å	Berndt, Güntert, Wüthrich (1993)	Acrescentam que o ângulo entre doador e aceptor deve ser menor que 35°
2,5 Å	Baker, Hubbard (1984)	Afirmam que o ângulo deve estar entre ±90 e 180°
0,80 – 3,49 Å	Taylor, Kennard e Versichel (1984)	
2,9 Å	Kabsch, Sander (1983)	

Distância	Publicação	Observações
2,65 – 3,39 Å	Wallwork (1962)	Em seu trabalho, Wallwork explora um grupo grande de ligações de hidrogênio, gerando os valores médios com seus respectivos desvios-padrão calculados.
2,26 – 3,36 Å	Nakamoto, Margoshes, Rundle (1955)	Em seu trabalho, levantam as distâncias para as pontes de hidrogênio baseando-se em várias referências da Literatura e considerando os átomos envolvidos como doadores e aceptores de próton, avaliando 57 distâncias diferentes.

As considerações sobre as distâncias que caracterizam as interações eletrostáticas, observadas na Literatura, estão na Tabela 7. Optou-se pela distância de 6 Å para caracterizar esta interação pois trata-se de valor que permitirá reduzir as possíveis falsas interações além de garantir que somente interações eletrostáticas mais fortes e mais precisas participarão da geração da base de dados.

Tabela 7 - Caracterização da distância da interação eletrostática

Distância	Publicação	Observações
6 Å	Harris et al (2011)	
3 – 5 Å	Takahashi (1997)	
7 Å	Lee, Warshel (1997)	Apesar de avaliarem valores 6, 7 e 8 Å, indicam como mais provável o valor 7 para a distância.

Os átomos que formam as interações aqui relacionadas estão descritos na Tabela 4, bem como o valor adotado para a distância e a quantidade de interações inseridas no banco de dados. Os valores escolhidos para a busca das distâncias, aqui definidos como *cutoff*, são parâmetros informados na execução dos *scripts* para a busca das interações para a geração dos arquivos e, portanto, passíveis de alteração implicando em aumento ou redução da quantidade de arquivos gerados com o aumento ou a redução, respectivamente, desse valor.

5.1.1 Etapa 1 – Montagem da base de dados inicial

Primeiramente, a base de dados das proteínas extraída do PDB (67.764 arquivos – provenientes da atualização do PDB ocorrida em setembro de 2010) foi percorrida a fim de se identificar os métodos de resolução de estrutura presentes no banco bem como a característica das interações existentes, resolução dos arquivos e mapeamento de todos os átomos, informações estas descritas nos arquivos PDB.

Pelo Gráfico 2 pode-se perceber que o maior número de arquivos foi resolvido com o método Difração de Raios-X (*X-Ray Diffraction*), 58.746 arquivos. Este método foi, portanto, escolhido para ter seus arquivos explorados, além disso, trata-se de um método que possui melhor confiabilidade das informações por possuir o valor da medida da resolução, medida que foi utilizada para resolver a estrutura, e também por ser um método que resolve a estrutura de um cristal, ou seja, uma estrutura com uma melhor precisão, permitindo que sejam descritos inclusive loops ou outras estruturas flexíveis não resolvidas por outros métodos experimentais. Dessa forma, o método e a resolução definidos no arquivo PDB foram os primeiros parâmetros para a escolha dos arquivos a serem utilizados neste projeto.

O Gráfico 3 apresenta o histograma das resoluções anotadas nos arquivos (registro "REMARK 3 RESOLUTION RANGE HIGH (ANGSTROMS) :") contidos na base de dados obtida no servidor de FTP[37] do PDB.

[37] Disponível em: ftp://ftp.wwpdb.org/

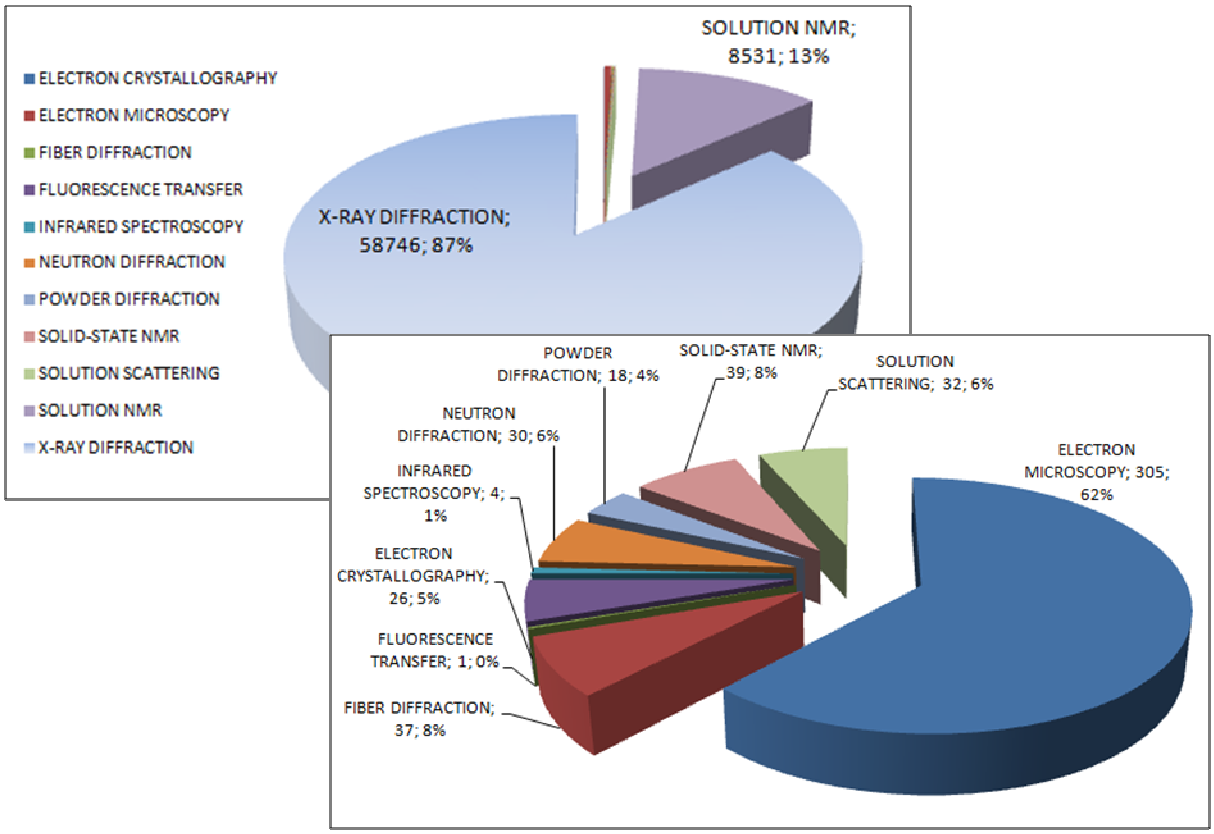


Gráfico 2 - Quantidade de métodos de resolução na base de dados: (Acima) Todos; (Abaixo) Excluindo X-Ray Diffraction (56799) e Solution NMR (8367)

Fonte: Dados extraídos dos arquivos do PDB (Setembro de 2010)

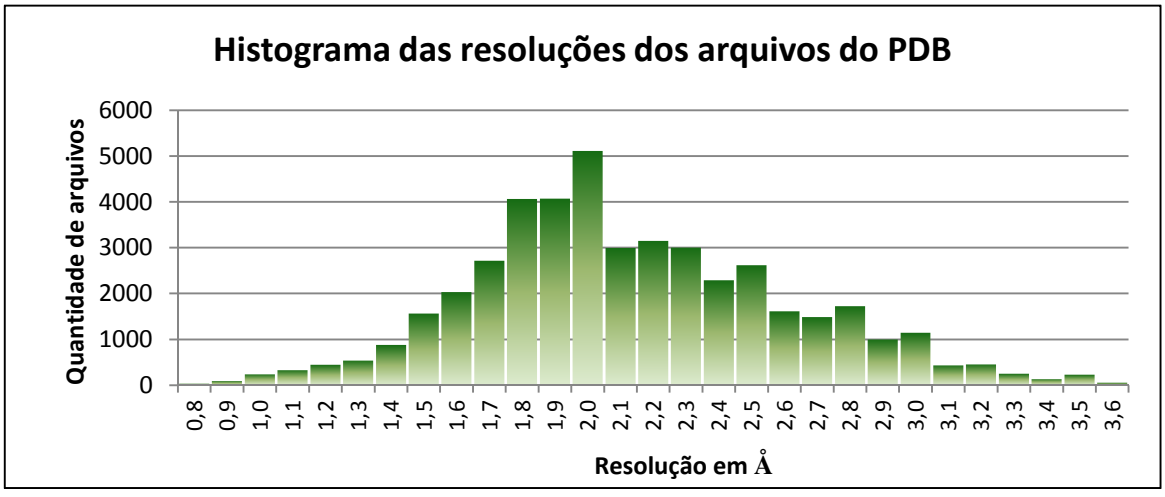


Gráfico 3 - Histograma das resoluções dos arquivos do PDB

Fonte: Dados extraídos dos arquivos do PDB (Setembro de 2010)

Para avaliação das pontes dissulfeto descritas nos arquivos PDB, o Gráfico 4 foi elaborado a partir das distâncias dessas pontes descritas nos arquivos PDB através do registro SSBOND.

Este gráfico foi gerado em duas escalas para demonstrar a grande concentração de arquivos em torno de 2,03 Å (escala direita).

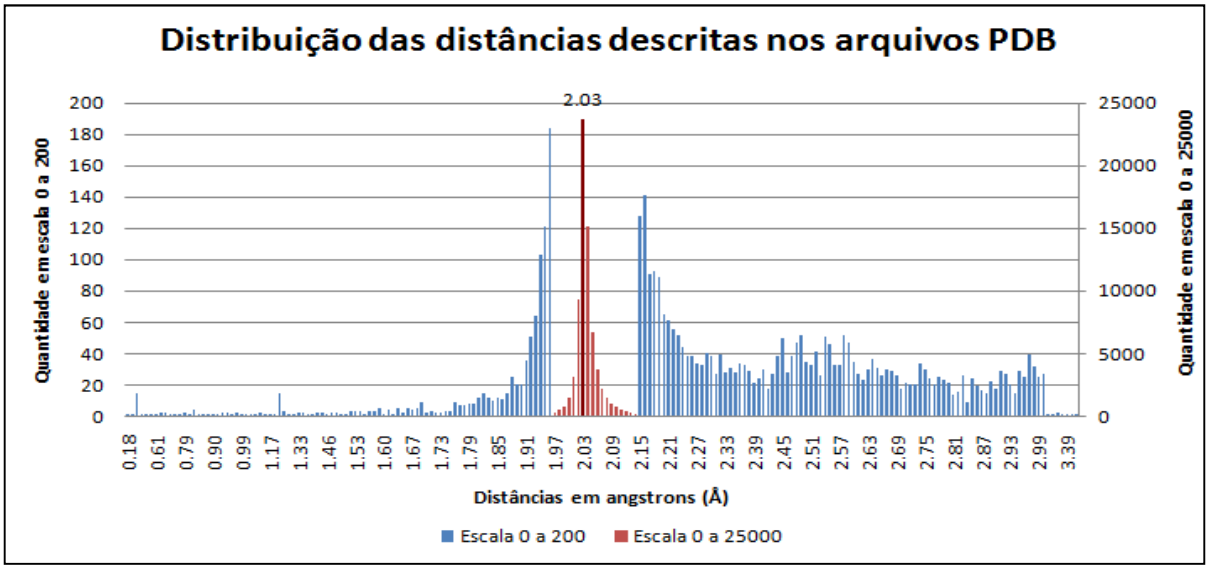


Gráfico 4 - Distribuição das distâncias S-S descritas nos arquivos PDB, em Angstrons

Fonte: Dados extraídos dos arquivos do PDB (Setembro de 2010)

O Gráfico 5 apresenta as mesmas distâncias do Gráfico 4, porém distribuídas de acordo com a resolução descrita no arquivo PDB. Percebe-se que a maior concentração das distâncias se dá em torno da resolução de 2 Å. Vale ressaltar que estas distâncias são anotadas nos arquivos PDB e por isso foram utilizadas na geração dos gráficos; as outras interações (ligação de hidrogênio e interação eletrostática) não possuem anotação de sua distância nos arquivos PDB para validação.

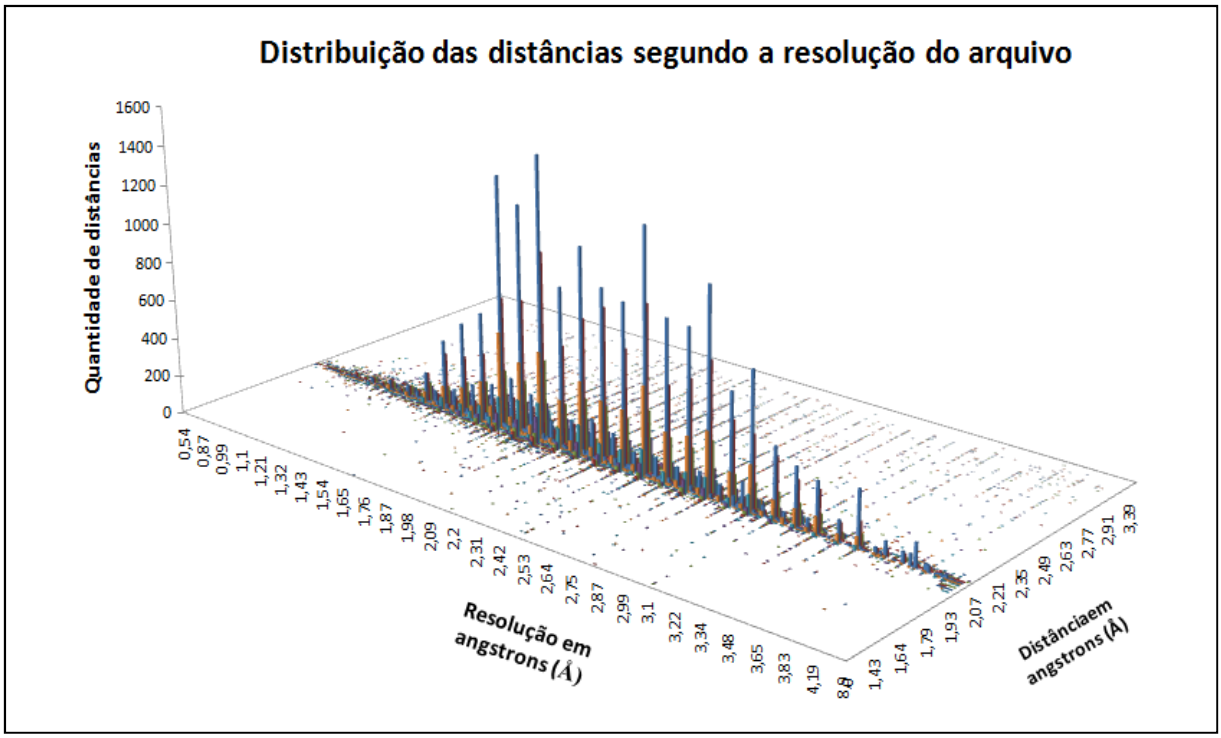


Gráfico 5 - Distribuição das distâncias S-S segundo a resolução do arquivo, em Angstrons

Fonte: Dados extraídos dos arquivos do PDB

Nesta etapa foram geradas três tabelas em nosso banco de dados:

- **methods** (11 registros) - tabela com todos os métodos descritos nos arquivos PDB (11 métodos) e a quantidade de arquivos de cada método presente na base de dados (Gráfico 2); o objetivo dessa tabela é agrupar os métodos similares e permitir a busca de arquivos por um método específico (no caso, X-Ray Diffraction);

- **pdbs** (67.764 registros) - tabela com os identificadores dos arquivos PDB (único para cada arquivo) bem como o seu método, a alta resolução e a baixa resolução. Estas duas resoluções dos arquivos definem a qualidade dos dados coletados, sendo que a alta resolução denota melhor confiabilidade da localização dos átomos na estrutura. Assim, quanto menor seu valor, melhor é a qualidade dos dados. O Gráfico 3 apresenta um histograma gerado a partir do levantamento obtido com esta tabela;

- **s_table** (820.972 registros) – tabela contendo todos os átomos de enxofre presentes em todos os arquivos PDB e todas as suas informações, a saber: identificador do arquivo PDB, número serial do átomo, identificação do átomo (exemplo carbono alfa,

oxigênio gama), nome do resíduo, número do resíduo, cadeia, coordenadas atômicas (x, y, z), dentre outras informações presentes na linha referente a cada átomo.

5.1.2 Etapa 2 – Cálculo das distâncias e definição dos parâmetros para caracterização das interações

Nesta etapa foi gerada a tabela distance_table (7.501.807 registros), contendo todas as distâncias calculadas entre os átomos das interações definidas de uma proteína, seguindo os critérios abaixo:

- A distância deve ser calculada através do método Euclidiano (Equação 5 anterior).

- Os átomos devem estar na mesma cadeia – o objetivo é identificar interações (pontes dissulfeto ou outra interação não covalente – ligação de hidrogênio, interação iônica, etc) numa mesma cadeia e não entre cadeias. Este outro tipo (entre cadeias) pode ser facilmente adicionado ao banco de interações posteriormente pela forma como o código foi desenvolvido. Sua exclusão neste momento permite reduzir a quantidade de variáveis a serem avaliadas no momento da validação e montagem da base de dados.

- As distâncias devem ser calculadas a partir de pares na direção N terminal para C terminal, ou seja, os pares de resíduos devem ter números crescentes, evitando-se assim duplicação de pontes (ou interações) contidas no banco com a inversão da ordem dos resíduos participantes.

Para análise dos arquivos e tabelas gerados, serão observados e relatados neste ponto os dados referentes à adição das pontes dissulfeto, primeira interação inserida no banco. Após a execução do módulo, dos 67.764 arquivos PDB, distribuídos em 11 diferentes métodos de resolução de estrutura, foram identificados 820.972 átomos de enxofre gama provenientes dos 58.746 arquivos PDB resolvidos por Difração de Raios-X. A partir desses átomos foram calculadas as distâncias entre eles, encontrando-se 7.501.807 pares de enxofre cuja distância foi calculada.

Dessas distâncias foram realizados filtros, descritos aqui, para se gerar os arquivos das interações que irão compor o banco. Assim, das distâncias calculadas, 111.953 estão abaixo de 3 Å e, mais precisamente, 105.554 são menores ou iguais a 2,4 Å. Se filtrarmos por átomos na mesma cadeia, encontramos 74.181 distâncias menores ou iguais a 3 Å e 71.493 menores ou iguais a 2,4 Å. Para garantir uma boa qualidade dos dados obtidos a partir desses arquivos, foi definida em 2 Å a resolução mínima do arquivo PDB de onde a ponte dissulfeto foi extraída. Assim, foram encontrados 60.614 pares de enxofres gama cujas distâncias são menores ou iguais a 2,4 Å, na mesma cadeia e com a resolução do arquivo PDB original maior ou igual a 2 Å (ou seja, o valor da resolução deve ser menor ou igual a 2 Å).

Estes 60.614 pares encontrados representam a base de dados inicial para as pontes dissulfeto encontradas no PDB. Estas pontes foram descritas em novos arquivos PDB gerados com apenas seis resíduos, ou seja, as cisteínas interagentes acrescidas dos resíduos anteriores e posteriores, com o objetivo de caracterizar a ponte com a menor quantidade de átomos possível, de forma a se obter melhores sobreposições na comparação com outros resíduos. Para caracterizar a cadeia principal da ponte são gerados arquivos específicos com os átomos da cadeia principal (N, Cα, C, O).

Porém, a geração dos arquivos teve que considerar a possibilidade de falta de resíduo imediatamente antes ou depois do resíduo alvo (cisteína no caso da ponte dissulfeto), o que pode caracterizar a falta de tal resíduo na anotação ou o resíduo alvo ser o primeiro ou o último resíduo da cadeia. Para identificar o resíduo anterior e posterior ao resíduo alvo a cada arquivo a ser gerado, foram calculadas as distâncias entre o carbono (C) do resíduo anterior e o nitrogênio (N) do resíduo alvo (resíduo central), bem como a distância entre o carbono (C) do resíduo alvo (resíduo central) e o nitrogênio (N) do resíduo posterior. Para validar a proximidade dos resíduos foi considerado o valor máximo de 1,6 Å para as distâncias C-N, valor este acima do descrito por Berg, Tymoczko e Stryer (2007), conforme Figura 19, 1,32 Å. Valores acima do determinado indicam que o resíduo real localizado próximo ao resíduo alvo (anterior ou posterior) não está descrito fielmente no arquivo (por problemas de anotação, resolução, dificuldade de identificação no mapa de densidade eletrônica) e que este resíduo não está próximo do alvo no espaço. Estes casos, quando percebidos, invalidam a geração do arquivo da possível ponte, pois o átomo de carbono do resíduo anterior e o nitrogênio do resíduo posterior ao resíduo alvo são necessários para o cálculo dos ângulos phi ϕ e psi ψ, necessários para determinar a conformação do resíduo central.

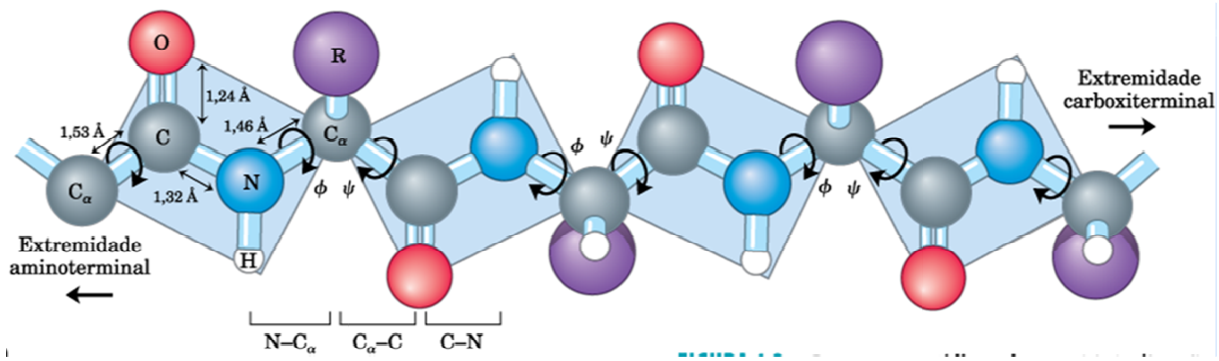


Figura 19 - Cadeia principal de uma proteína com seus comprimentos típicos e os ângulos phi ϕ e psi ψ

Fonte: Lehninger, Nelson, Cox (2007)

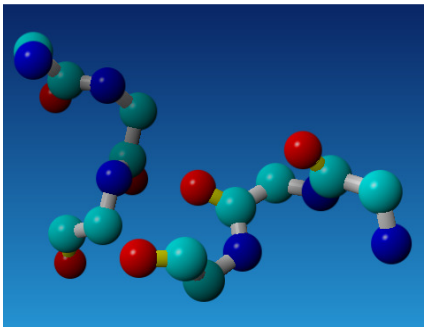
Dessa forma, as seguintes características descartam do banco o arquivo da interação gerado:

- Ausência de resíduo anterior ou posterior ao átomo observado – impossibilitando, portanto, que sejam calculados os ângulos phi ϕ e psi ψ.

- Ausência de algum átomo na cadeia principal.

- Erros de anotação, como os descritos na seção 5.1.6 (Problemas encontrados em arquivos do PDB), envolvendo: anotação incorrea de dupla conformação, anotação incorreta de alguns números como o número do átomo, número do resíduo, altloc, dentre outros - Estes erros ocasionam a identificação incorreta do resíduo pelos *scripts* gerados, resultando em distâncias, nomes ou sequências de resíduos incorretos.

Estes critérios levaram os 60.614 arquivos, relacionados às pontes dissulfeto, à eliminação de alguns que não atendiam aos requisitos acima, chegando-se à quantidade de 48.523 arquivos, contendo 6 resíduos cada como mostrado, em um exemplo, na Figura 20, a seguir, que contém o arquivo PDB gerado e a imagem tridimensional dos resíduos. A numeração original dos resíduos de aminoácidos é alterada pelo *script* que gera o arquivo, com o objetivo de identificar os resíduos (1, 2 e 3 – um lado; 101, 102, 103 – outro lado) durante a execução da sobreposição por outro *script*.



```
CRYST1   73.678   41.460  119.863  90.00 103.56  90.00 P 1 21 1     4
SCALE1      0.013573  0.000000  0.003274        0.00000
SCALE2      0.000000  0.024120  0.000000        0.00000
SCALE3      0.000000  0.000000  0.008582        0.00000
ATOM      1  N   LEU A   1      10.555   9.791   7.655  1.00 32.19      N
ATOM      2  CA  LEU A   1      11.780   9.222   7.093  1.00 31.25      C
ATOM      3  C   LEU A   1      12.969   9.585   7.949  1.00 31.45      C
ATOM      4  O   LEU A   1      13.837   8.765   8.132  1.00 30.25      O
ATOM      5  N   CYS A   2      13.019  10.815   8.446  1.00 33.15      N
ATOM      6  CA  CYS A   2      14.069  11.212   9.391  1.00 33.81      C
ATOM      7  C   CYS A   2      13.991  10.400  10.682  1.00 32.76      C
ATOM      8  O   CYS A   2      15.006   9.982  11.224  1.00 33.63      O
ATOM      9  N   THR A   3      12.791  10.207  11.207  1.00 32.46      N
ATOM     10  CA  THR A   3      12.618   9.437  12.455  1.00 31.52      C
ATOM     11  C   THR A   3      13.106   7.989  12.320  1.00 31.43      C
ATOM     12  O   THR A   3      13.722   7.424  13.257  1.00 30.48      O
ATOM     13  N   LYS A 101      20.435   8.227  12.942  1.00 36.62      N
ATOM     14  CA  LYS A 101      21.226   9.465  13.034  1.00 36.28      C
ATOM     15  C   LYS A 101      20.384  10.731  12.886  1.00 34.37      C
ATOM     16  O   LYS A 101      20.565  11.691  13.623  1.00 31.45      O
ATOM     17  N   CYS A 102      19.462  10.750  11.908  1.00 32.75      N
ATOM     18  CA  CYS A 102      18.616  11.907  11.700  1.00 32.77      C
ATOM     19  C   CYS A 102      17.619  12.116  12.843  1.00 32.04      C
ATOM     20  O   CYS A 102      17.362  13.259  13.252  1.00 33.05      O
ATOM     21  N   ASP A 103      17.035  11.017  13.322  1.00 31.07      N
ATOM     22  CA  ASP A 103      16.070  11.089  14.405  1.00 31.54      C
ATOM     23  C   ASP A 103      16.716  11.755  15.652  1.00 32.07      C
ATOM     24  O   ASP A 103      16.150  12.647  16.278  1.00 32.62      O
END
```

Figura 20 - Arquivo SG2gh0_169B_175B.ent-f.trans-mc.pdb, contendo a cadeia principal da ponte CYS169-CYS175, da cadeia B, do arquivo pdb2gh0.ent.

Segundo Cohen, Potapov, Schreiber (2009), um conjunto de quatro distâncias entre dois resíduos interagentes (Figura 21) pode caracterizar uma interação. Teoricamente, segundo os autores, para uma descrição mais precisa e completa da interação seriam necessárias múltiplas distâncias entre os dois resíduos.

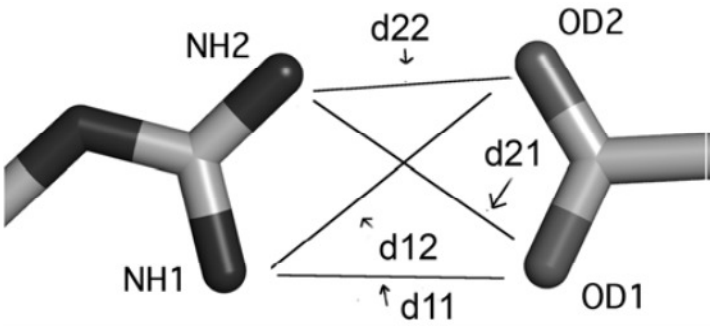


Figura 21 - As quatro distâncias para o par de resíduos interagentes Arg-Asp
Fonte: Cohen, Potapov, Schreiber (2009) - doi:10.1371/journal.pcbi.1000470.g001

Os arquivos gerados para cada interação (exemplo na Figura 20) foram percorridos e foram calculadas 16 distâncias compreendendo cada átomo da cadeia principal (N, Cα, C, O) do primeiro resíduo alvo (Cys-2) e estes mesmos átomos do segundo resíduo alvo (Cys-102) , que formam a interação. O objetivo é utilizar os limites definidos pelas distâncias encontradas para identificar, na proteína alvo, possíveis conformações candidatas a terem sua cadeia lateral mutada para o resíduo alvo em questão (cisteína, no caso da ponte dissulfeto, ou outro

conforme a interação proposta). Os dois gráficos a seguir apresentam a distribuição dessas distâncias para o Cα e o C, para o caso das pontes dissulfeto. Estas distâncias calculadas foram diferentes para cada interação inserida no banco de dados. O Gráfico 6 refere-se às distâncias do átomo Cα de um resíduo aos átomos N, Cα, C e O do outro resíduo, enquanto que o Gráfico 7 apresenta as distâncias do átomo C. O valor mais incidente observado para a distância Cα-Cα, 5,8 Å está de acordo com a faixa descrita por Richardson (1981), 4 Å a 7,4 Å, e menor que 6,5 Å, como citado por Sowdhamini *et al* (1989). Estes dois autores abordam a distância Cα-Cα como uma característica relevante para se identificar uma ponte dissulfeto.

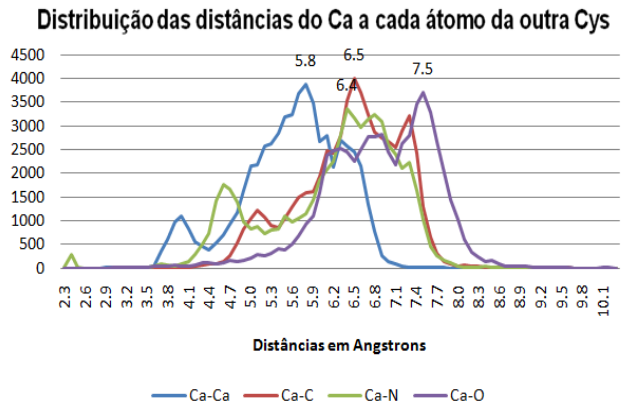


Gráfico 6 - Distribuição das distâncias, em Angstrons, do Cα (CA) a cada átomo da outra Cys
Fonte: Dados extraídos dos arquivos do PDB

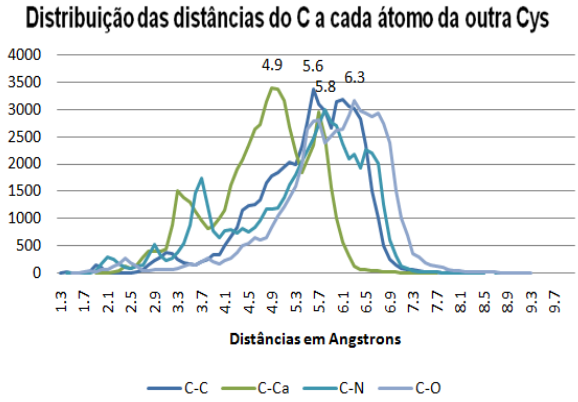


Gráfico 7 - Distribuição das distâncias, em Angstrons, do C a cada átomo da outra Cys
Fonte: Dados extraídos dos arquivos do PDB

Após a geração dos arquivos do banco, estes foram, um a um, sobrepostos ao arquivo de maior resolução contido no banco (PDB 1EJG[38], resolução 0,540 Å, distância da ponte dissulfeto de 2,03 Å). Esta sobreposição resulta num arquivo transformado por rotação e translação das coordenadas atômicas na mesma orientação do arquivo de maior resolução, com o objetivo futuro de realizar comparação direta dos arquivos, ou seja, arquivos que apesar de apresentarem a interação em posições espaciais diferentes estão relacionadas por uma rotação e uma translação simples.

[38] PDB ID 1EJG. Jelsch, C., Teeter, M.M., Lamzin, V., Pichon-Pesme, V., Blessing, R.H., Lecomte, C. Accurate protein crystallography at ultra-high resolution: valence electron distribution in crambin. Journal: (2000) Proc.Natl.Acad.Sci.USA 97: 3171-3176. PubMed: 10737790. PubMedCentral: PMC16211.

5.1.3 Etapa 3 – Busca em uma proteína alvo

Para esta etapa foi desenvolvido o módulo SG-search que percorre a proteína alvo em busca de pares de resíduos passíveis de serem mutados a fim de se inserir naquele ponto uma interação. Como exemplo será demonstrada uma busca realizada no banco de dados de pontes dissulfeto.

Com o objetivo de testar o algoritmo desenvolvido, simulando uma busca em uma proteína alvo de estrutura recém-resolvida, na qual se pretende propor mutações para ponte dissulfeto, foi escolhido um polipeptídeo com estrutura resolvida, depositado no PDB. Portanto, como proteína alvo, foi escolhido o arquivo PDB 1PEN[39], Figura 22.

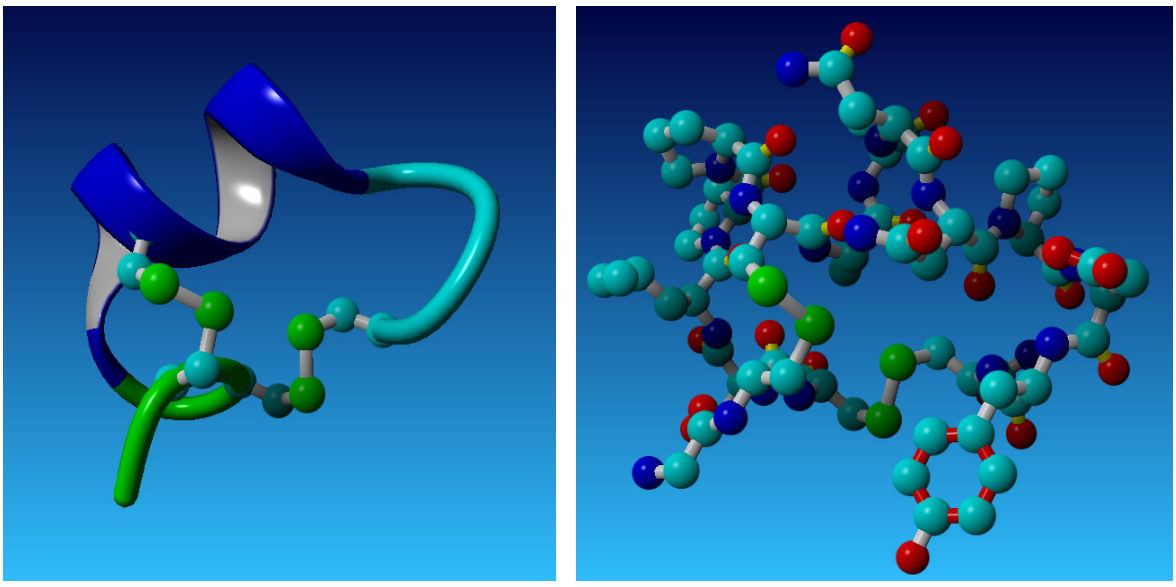


Figura 22 - Estrutura do arquivo 1pen, demonstrando alfa hélices, loops e pontes dissulfeto (átomos em verde destacados à esquerda). À direita, sua estrutura atômica completa.

Esta escolha se deu pelo fato de este arquivo possuir duas pontes dissulfeto anotadas, Figura 23, e por possuir uma cadeia de 109 átomos (num total de 16 resíduos em uma única cadeia), bom tamanho para testar o algoritmo completo (percorrendo toda a base com 48.523 arquivos; alguns redundantes). As duas pontes presentes (CYS2A-CYS8A e CYS3A-CYS16A) possuem, ambas, distância de 2.03 Å entre os átomos de enxofre gama (SG). O arquivo apresenta resolução de 1,1 Å e sua estrutura foi resolvida por Difração de Raios-X.

[39] PDB ID 1PEN. Hu, S.H., Gehrmann, J., Guddat, L.W., Alewood, P.F., Craik, D.J., Martin, J.L. The 1.1 Å crystal structure of the neuronal acetylcholine receptor antagonist, alpha-conotoxin PnIA from Conus pennaceus, PubMed 8740364, Structure. 1996 Apr 15;4(4):417-23.

```
SSBOND   1 CYS A    2    CYS A    8                      1555   1555  2.03
SSBOND   2 CYS A    3    CYS A   16                      1555   1555  2.03
```

Figura 23 - Trecho SSBOND do arquivo pdb1pen.ent, descrevendo as pontes dissulfeto

A razão de se escolher um polipeptídeo já com pontes dissulfeto é a possibilidade de validar estas pontes existentes na busca realizada no banco de dados, ou seja, se removermos a ponte do polipeptídeo, podemos avaliar se ela seria encontrada. Se a proteína alvo pertencer ao banco, pode-se remover do banco as pontes encontradas e verificar se aqueles resíduos são identificados como candidatos a mutação.

A primeira parte dessa etapa consistiu em percorrer a cadeia principal do polipeptídeo em busca de pares de resíduos que atendam às faixas de distâncias definidas na etapa anterior. Estas distâncias foram coletadas de todos os arquivos que descrevem pontes a fim de se definir padrões ou limites máximo e mínimo das distâncias entre os quatro átomos da cadeira principal, num total de 16 distâncias, apresentados no trecho inicial da Figura 24. Sendo assim, a cada par de resíduos da proteína alvo, quatro dessas distâncias foram calculadas (N-N, Cα-Cα, C-C, O-O) e, se estiver dentro do padrão levantado, este par passa a fazer parte dos pares de resíduos da proteína que são passíveis de terem sua cadeia lateral mutada para cisteína.

```
log> 27/03/2011 - 03:48:28
log> Directory that will be used: 0218-distances-201103271541-ok
Distances CACA max = 8.962; min = 2.898
Distances CAC max = 9.745; min = 3.454
Distances CAN max = 9.478; min = 2.315
Distances CAO max = 10.322; min = 2.713
Distances CCA max = 9.724; min = 2.342
Distances CC max = 11.087; min = 2.951
Distances CN max = 10.494; min = 1.304
Distances CO max = 11.674; min = 2.717
Distances NCA max = 9.763; min = 3.115
Distances NC max = 10.610; min = 3.525
Distances NN max = 10.307; min = 2.582
Distances NO max = 11.499; min = 2.549
Distances OCA max = 10.488; min = 2.659
Distances OC max = 11.785; min = 2.773
Distances ON max = 11.227; min = 2.118
Distances OO max = 12.359; min = 2.492

Residue 1: GLY  1  A
Residue 2: CYS  2  A

Residue 1: GLY  1  A
Residue 2: CYS  3  A

Residue 1: GLY  1  A
Residue 2: SER  4  A

Residue 1: GLY  1  A
Residue 2: LEU  5  A

Residue 1: GLY  1  A
Residue 2: CYS  8  A

Residue 1: CYS  2  A
Residue 2: CYS  3  A
```

Figura 24 - Trecho do relatório do módulo SG-search, que percorre a proteína identificando pares de resíduos

A Figura 24 apresenta a relação de valores de distâncias gerados pelo relatório. Os valores médios se encontram de acordo com a Literatura e a grande variação na distribuição dos valores está diretamente relacionada às variações nas conformações das cadeias principais dos resíduos interagentes. As linhas seguintes às distâncias mostradas na Figura 24 apresentam os primeiros 6 dos 58 pares de resíduos candidatos encontrados no polipeptídeo, sendo o primeiro par da lista composto pela Glicina (resíduo de número 1 da cadeia A) e a Cisteína (resíduo de número 2 da cadeia A).

Para cada um dos 58 pares encontrados foi gerado um arquivo no formato pdb, idêntico aos arquivos gerados para o banco, ou seja, contendo apenas os resíduos que compõem o par e os resíduos anteriores e posteriores a estes, além de alguns cabeçalhos necessários para a rotação e translação dos átomos. Estes arquivos foram verificados e os que não estavam de acordo com o padrão gerado foram eliminados. Os arquivos eliminados não continham um dos resíduos (anterior ou posterior), ou então não continham algum átomo da cadeia principal, ou

continham erros de anotação, descritos na seção 5.1.6 (Problemas encontrados em arquivos do PDB). Dessa forma, restaram 41 arquivos com os pares de resíduos candidatos à mutação por cisteína da proteína-alvo. Estes arquivos foram, um a um, sobrepostos aos arquivos do banco usando o aplicativo LSQKAB do pacote CCP4[40]. Esta sobreposição visa comparar as conformações da cadeia principal dos pares de resíduos (alvo e banco), gerando 4 arquivos de saída, sendo:

1. **Arquivo ".out"** – resultado da transformação contendo as novas coordenadas para os arquivos do banco geradas após sobreposição dos mesmos contra cada arquivo da proteína-alvo.

2. **Arquivo ".rmstab"** – apresenta uma distância média por resíduo obtida a partir das distâncias computadas entre os pares de átomos equivalentes das cadeias principais de cada resíduo após a sobreposição.

3. **Arquivo ".log"** – contem um grupo de mensagens geradas na execução do script, para verificação de erros que porventura possam ter ocorrido durante a execução. Apresenta também os vetores que foram usados para a rotação e translação da molécula resultante.

4. **Arquivo ".deltas"** – apresenta a distância entre os pares de átomos equivalentes das cadeias principais de cada resíduo após a sobreposição dos dois arquivos, conforme Figura 28. Vale ressaltar que apenas os átomos de carbono C e nitrogênio N da cadeia principal do primeiro e do terceiro resíduos de aminoácidos, respectivamente, bem como todos os átomos da cadeia principal do segundo resíduo foram usados na sobreposição. O objetivo dessa estratégia é poder sobrepor exatamente o resíduo que será mutado (resíduo central) mais um átomo imediatamente anterior e um imediatamente posterior, pois este conjunto é responsável pelo cálculo dos ângulos phi e psi. Esta quantidade de átomos usada na sobreposição pode ser alterada para todos os átomos da cadeia principal, para que se tenha uma sobreposição mais restritiva e precisa dos três resíduos consecutivos.

[40] Collaborative Computational Project No. 4. Software for Macromolecular X-Ray Crystallography. Disponível em http://www.ccp4.ac.uk

Ao todo foram gerados, para a sobreposição de 41 arquivos do alvo contra os 48.523 arquivos do banco, 1.989.443 grupos de 4 arquivos totalizando 39,4 GB (deltas – 7.8 GB, log – 16 GB, out – 7.8 GB e rmstab – 7.8 GB), consumindo 24 horas e 10 minutos de processamento, o que justifica a necessidade de otimização do procedimento.

A sequência de pares de resíduos gerada pela ferramenta desenvolvida está listada na Figura 25, a seguir. Esta sequência foi obtida a partir da verificação das distâncias entre os átomos de todos os pares possíveis do polipeptídeo (proteína alvo) em comparação com os valores dessas distâncias extraídos do pdb. Portanto, são pares que ainda devem ser analisados para se propor a mutação.

ALA9A – ALA10A	CYS2A – ASN12A	CYS3A – PRO6A	PRO6A – ALA9A
ALA9A – ASN12A	CYS2A – CYS3A	CYS3A – SER4A	PRO6A – PRO7A
ALA9A – PRO13A	CYS2A – CYS8A	CYS3A – TYR15A	PRO7A – ALA10A
ALA9A – TYR15A	CYS2A – LEU5A	CYS8A – ALA9A	PRO7A – ASN11A
ALA10A – ASN11A	CYS2A – PRO6A	CYS8A – ASN11A	PRO7A – ASN12A
ALA10A – PRO13A	CYS2A – PRO7A	CYS8A – ASN12A	PRO7A – CYS8A
ASN11A – ASN12A	CYS3A – ALA10A	LEU5A – ALA9A	PRO13A – ASP14A
ASN12A – PRO13A	CYS3A – ALA9A	LEU5A – CYS8A	SER4A – ALA9A
ASN12A – TYR15A	CYS3A – ASN12A	LEU5A – PRO6A	SER4A – CYS8A
ASP14A – TYR15A	CYS3A – CYS8A	PRO6A – ALA10A	SER4A – LEU5A
CYS2A – ALA9A			

Figura 25 - Lista dos pares candidatos a mutação encontrados ao término da execução do módulo

Cada par listado foi sobreposto com todas as pontes do banco (48.523), resultando em vários arquivos com o resultado da sobreposição e deltas, conforme descrito anteriormente. Desses arquivos, pode-se gerar a Tabela 8, que apresenta as maiores distâncias interatômicas encontradas (entre os 12 átomos equivalentes) como resultado da sobreposição entre os 48.523 arquivos PDBs oriundos do banco de dados e os 41 arquivos PDBs provenientes da proteína alvo. Ela foi gerada com o objetivo de se observar os menores valores dessa lista que podem significar uma sobreposição muito boa quando o valor estiver muito próximo de zero, permitindo inferir possíveis clusterizações (agrupamentos) entre as sobreposições. A primeira linha (0,001 Å) apresenta a sobreposição do par CYS2A-CYS8A (ponte dissulfeto presente no polipeptídeo) com seu similar existente no banco de dados. Este mesmo par da proteína alvo também apresenta valores baixos na sobreposição com outras pontes do banco, apresentadas nas linhas 2 (0,163 Å), 3 (0,168 Å), 4 (0,203 Å) e 6 (0,290 Å), dentre outras, o que sugere um agrupamento (cluster) dessas pontes no banco como similares. A outra ponte do polipeptídeo

(CYS3A-CYS16A) não figura no relatório (Figura 25) e tampouco na listagem (Tabela 8) pois como um dos resíduos é o último da cadeia, este arquivo não foi gerado.

As linhas 5 e 15, da Tabela 8, referem-se à sobreposição do par ASP14A-TYR15A da proteína alvo com valores baixos para a maior distância interatômica (0,270 e 0,340 Å). Isto indica que este par é um candidato à mutação das cadeias laterais do primeiro resíduo (aspartato) e do segundo resíduo (tirosina) para cisteínas, possibilitando naquele ponto a formação de uma ponte dissulfeto, pois este par e as duas pontes do banco de dados indicadas na tabela apresentam conformações muito próximas, com uma diferença máxima de 0,270 Å ou 0,340 Å em um dos átomos.

Tabela 8 - Maior distância interatômica por arquivo

	Maior distância (Å)	PDB alvo	1º Resíduo	2º Resíduo	PDB banco	CYS1	CYS2
1	0.001	1pen	CYS2A	CYS8A	1pen	2A	8A
2	0.163	1pen	CYS2A	CYS8A	1akg	2A	8A
3	0.168	1pen	CYS2A	CYS8A	2dqa	59B	65B
4	0.203	1pen	CYS2A	CYS8A	3fub	161C	167C
5	0.270	1pen	ASP14A	TYR15A	3gtk	45J	46J
6	0.290	1pen	CYS2A	CYS8A	2gh0	169B	175B
7	0.310	1pen	CYS2A	CYS8A	2v5e	161A	167A
8	0.315	1pen	ASN11A	ASN12A	2d0v	103D	104D
9	0.323	1pen	CYS3A	SER4A	2e2i	45J	46J
10	0.326	1pen	CYS2A	CYS8A	2uz6	2N	8N
11	0.333	1pen	CYS2A	CYS8A	2v5e	250A	256A
12	0.334	1pen	CYS2A	CYS8A	2gh0	169A	175A
13	0.335	1pen	LEU5A	CYS8A	2hj3	54A	57A
14	0.339	1pen	CYS2A	CYS8A	2uz6	2T	8T
15	0.340	1pen	ASP14A	TYR15A	2nvz	45J	46J
16	0.347	1pen	CYS2A	CYS8A	1a0m	2B	8B
17	...	...	...	...	...	...	...
18	0.380	1pen	ASN12A	TYR15A	1gai	210A	213A

Outros pares da tabela apresentam conformações similares às presentes no banco de conformações com um valor baixo para a distância interatômica, a saber, ASN11A-ASN12A, CYS3A-SER4A e LEU5A-CYS8A. Estes dois últimos afetam as pontes existentes que envolvem as cisteínas 2A-8A e 3A-16A. Além desses, há vários outros resultados (1.989.443

resultados) não exibidos na Tabela 8 (razão da representação "..." na linha 17 da tabela). Interessante observar o par da linha 18 (ASN12A-TYR15A) que difere dos demais por não conter cisteína e não serem aminoácidos consecutivos, o que possibilita a manutenção da estabilidade neste ponto do polipeptídeo (Figura 26 e Figura 27).

A Figura 26 apresenta duas sobreposições de pontes dissulfeto do banco sobre o polipeptídeo. A primeira (à esquerda da figura) refere-se à ponte CYS2B-CYS8B do arquivo PDB 1a0m (linha 16 da Tabela 8), em cinza, o que demonstra que, caso o arquivo 1pen fosse removido do banco, esta ponte seria encontrada, validando o algoritmo de busca. Observa-se pela Tabela 8 a existência de várias outras pontes (linhas 2, 3, 4, 6, 7, 10, 11, 12, 14) que apresentam distâncias pequenas quando sobrepostas à ponte já existente no polipeptídeo (CYS2A-CYS8A). Por outro lado, à direita da Figura 26, encontra-se a ponte CYS10A-CYS13A do arquivo PDB 1gai (linha 18 da Tabela 8) em amarelo. Esta sobreposição permite indicar esta ponte como uma possível mutação no polipeptídeo alvo 1pen, permitindo a introdução de uma nova ponte dissulfeto em sua estrutura, como visto na Figura 27.

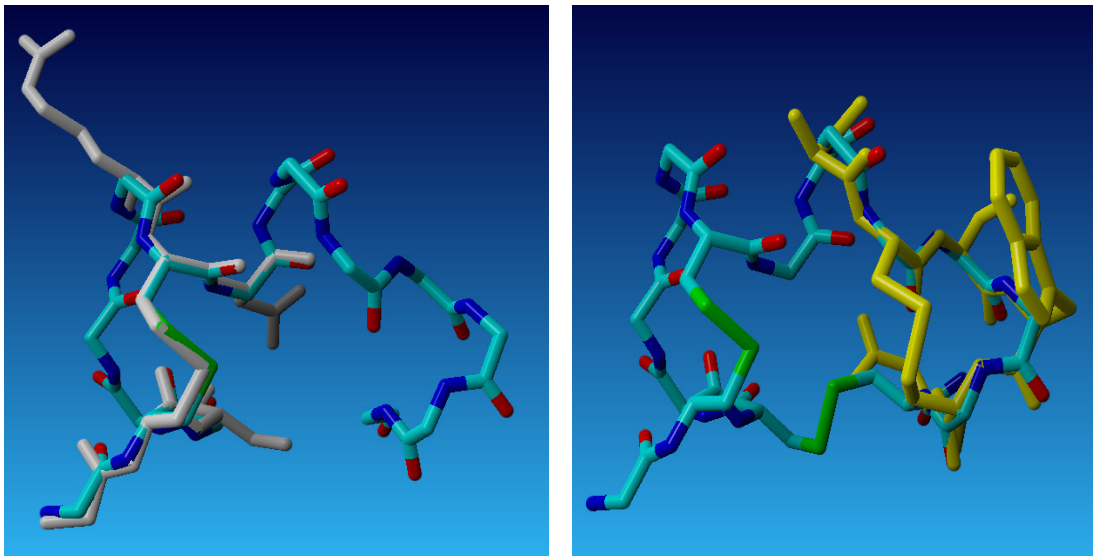


Figura 26 - 1pen com as sobreposições da linha 16 da Tabela 8 (ponte CYS2B-CYS8B do pdb 1a0m) à esquerda e da linha 18 (ponte CYS210A-CYS213A do pdb 1gai)

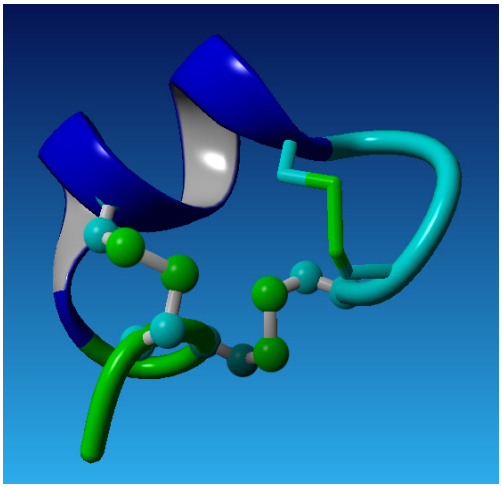


Figura 27 – Estrutura do polipeptídeo PDB 1PEN e suas pontes dissulfeto (átomos de enxofre em verde) além de uma possível ponte a ser adicionada a partir do banco (1gai – CYS210A-CYS213A) à direita da figura

A Figura 28, a seguir, apresenta dois arquivos ".deltas" gerados na sobreposição anterior. O arquivo gerado possui um nome composto por duas partes. A primeira parte se refere ao nome da proteína alvo, ao par de resíduos da proteína alvo , seus números e sua cadeia. A segunda parte se refere ao nome do arquivo do banco de dados de onde as interações foram retiradas, os números das cisteínas (no caso da busca por pontes dissulfeto), seguidos da cadeia. O primeiro arquivo de resultado (pdb1pen_CYS2A_CYS8A-SG1pen_2A_8A.deltas) refere-se à sobreposição do arquivo do par proveniente do peptídeo alvo (pdb1pen_CYS2A_CYS8A) com o arquivo da ponte existente no banco (SG1pen_2A_8A) também referente ao mesmo par do polipeptídeo, contido no banco. Sua análise permite confirmar a sobreposição perfeita (maior distância interatômica igual a 0,001 Å) átomo a átomo da ponte do polipeptídeo com a mesma existente no banco (esquerda da figura).

A outra sobreposição relatada na figura (direita – pdb1pen_ASN12A_TYR15A-SG1gai_210A_213A.deltas) refere-se à sobreposição de um par do polipeptídeo alvo com uma ponte do banco indicando que este par é um candidato à mutação da cadeia lateral uma vez que o maior valor de distância interatômica indicado na lista é 0,380 Å. Este valor muito próximo de zero indica que os dois pares sobrepostos se encontram em posições muito próximas com uma diferença de apenas 0,380 Å entre os átomos de N do resíduo posterior à segunda cisteína da ligação. Os outros valores, correspondentes aos outros átomos, são menores que esse conforme se observa na Figura 28 (direita). Esta sobreposição poderá ser observada na Figura 26.

pdb1pen_CYS2A_CYS8A-SG1pen_2A_8A.deltas					pdb1pen_ASN12A_TYR15A-SG210A_213A.deltas				
0.001	1C	A	1C	A	0.283	1C	A	1C	A
0.001	2N	A	2N	A	0.230	2N	A	2N	A
0.001	2CA	A	2CA	A	0.280	2CA	A	2CA	A
0.001	2C	A	2C	A	0.143	2C	A	2C	A
0.001	2O	A	2O	A	0.076	2O	A	2O	A
0.000	3N	A	3N	A	0.094	3N	A	3N	A
0.000	101C	A	101C	A	0.319	101C	A	101C	A
0.000	102N	A	102N	A	0.150	102N	A	102N	A
0.000	102CA	A	102CA	A	0.348	102CA	A	102CA	A
0.000	102C	A	102C	A	0.066	102C	A	102C	A
0.001	102O	A	102O	A	0.267	102O	A	102O	A
0.000	103N	A	103N	A	0.380	103N	A	103N	A

Figura 28 - Arquivos deltas para a ponte 2A-8A (esquerda) e outro par do banco ASP14A-TYR15A (direita)

A Figura 29, a seguir, gerada utilizando-se a ferramenta Yasara, mostra a sobreposição dos arquivos referentes às linhas 1, 2, 3, 4 e 6 da Tabela 8, que se sobrepõe à ponte CYS2A_CYS8A do polipeptídeo, também encontrada no banco (pois a mesma faz parte da base) e as demais que se apresentaram bastante similares conformacionalmente. A figura apresenta apenas os átomos da cadeia principal (N – azul, C e Cα – ciano, O – vermelho). Sua observação permite concluir o quão próximas estão as conformações das cadeias principais dos arquivos sobrepostos.

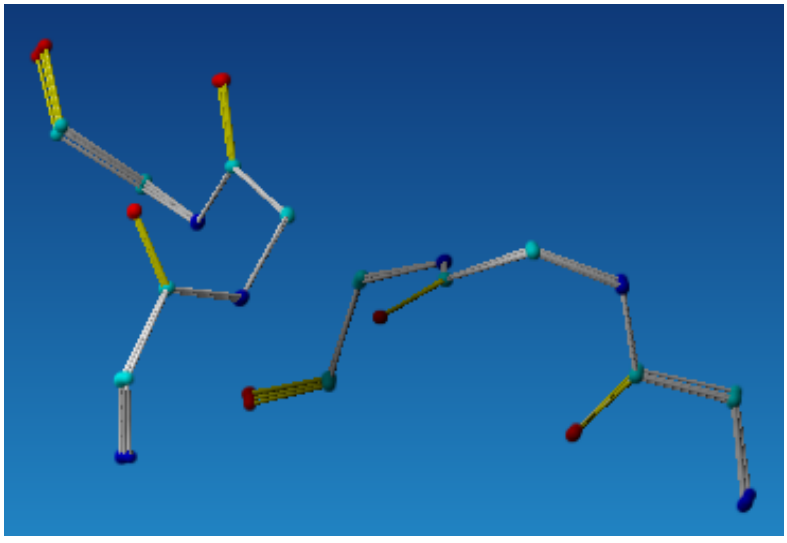


Figura 29 - Exibição das linhas 1, 2, 3, 4 e 6 da Tabela 8 (pontes do banco que se sobrepõem com menores distâncias à ponte CYS2A-CYS8A do polipeptídeo 1pen)

5.1.3.1 Comparação do resultado obtido com o resultado de outras ferramentas

A partir da sequência de resíduos do polipeptídeo 1PEN (sequência fasta: GCCSLPPCAANNPDYCX) vários resultados foram obtidos em ferramentas de predição de ponte dissulfeto, conforme análise que se segue:

- ✓ EDBCP (Figura 30): Prediz as pontes já existentes (2-8, 3-16), com os maiores valores de probabilidade (0,65310 e 0,41345, respectivamente), além de avaliar outras possíveis ligações entre as cisteínas existentes (2-3, 2-16, 3-8, 8-16). Excetuando-se os pares que contém a cisteína 16 (por ser o último resíduo da cadeia), os outros foram também propostos pela ferramenta apresentada neste trabalho.

Prediction Result

Job ID	20121209289	Query Name	1PEN
Template PDB ID	1PENA		
SEQUENCE	GCCSLPPCAANNPDYCX Sequence length: 17 residues		
Positions of cysteines	2 3 8 16		
e-value	0.005	Sequence identity	100.00%

Disulfide bond connectivity prediction score

Position	Cysteines pair Euclidean distance	Separation	Probability	Metal-binding site score
2 - 3	3.76	1	0.01194	0.21990
2 - 8	4.79	6	0.65310	0.21990
2 - 16	7.38	14	0.14301	0.21990
3 - 8	7.10	5	0.09184	0.18023
3 - 16	10.34	13	0.41345	0.21989
8 - 16	7.37	8	0.05618	0.23607
Positions of oxidized cysteines	2 3 8 16			
Predicted disulfide bonds	[2-8] [3-16]			
Predicted positions of cysteines in metal binding site	None			

Figura 30 - Trecho do resultado da execução do EDBCP

- ✓ DiANNA (Figura 31): Ao invés das pontes existentes (2-8, 3-16) prediz como pontes os pares 3-8 (maior *score*, 0,01201) e 2-16 (*score* 0,01042, maior *score* que não envolve as cisteínas 3 e 8). Também avalia as possíveis ligações entre as cisteínas existentes em pares como 2-3, 2-8, 3-16 e 8-16.

Step 4: Disulfide Bonds Prediction using a trained Neural Network

Disulfide bond scores

Cysteine sequence position	Distance	Bond	Score
2 - 3	1	XXXXGCCSLPP-XXXGCCSLPPC	0.01038
2 - 8	6	XXXXGCCSLPP-CSLPPCAANNP	0.01184
2 - 16	14	XXXXGCCSLPP-NNPDYCXXXXX	0.01042
3 - 8	5	XXXGCCSLPPC-CSLPPCAANNP	0.01201
3 - 16	13	XXXGCCSLPPC-NNPDYCXXXXX	0.01054
8 - 16	8	CSLPPCAANNP-NNPDYCXXXXX	0.01043

Step 5: Weighted matching

Predicted bonds	
2 - 16	XXXXGCCSLPP - NNPDYCXXXXX
3 - 8	XXXGCCSLPPC - CSLPPCAANNP

Predicted connectivity
1-4, 2-3

Figura 31 - Trecho do resultado da execução do DiANNA

✓ Disulfind (Figura 32): Lista as pontes já existentes (cisteínas 2-8 e 3-16).

Results for **1pen**

```
          +------------+
         +|----+       |
         ||    |       |
        .........10......
AA      GCCSLPPCAANNPDYCX
DB_state  11    1       1
DB_conf   78    8       7

DB_bond   bond(2,8)
DB_bond   bond(3,16)

Conn_conf 0.873012
```

Figura 32 - Trecho do resultado da execução do Disulfind

✓ SSBOND (Figura 33): Lista as pontes já existentes (2-8 e 3-16) além de mais um par LEU5A-CYS8A para possível ponte, baseados na distância entre os carbonos beta (CB) e posteriormente entre os enxofres gama (SG). Este par também é levantado pelo algoritmo aqui descrito (linha 13 da Tabela 8) com a maior distância interatômica igual a 0,335 Å.

```
   SSBOND   CHI1-1 CHI2-1   CHI3 CHI2-2 CHI1-2 SGDIST CHINRG TAUNRG DISNRG TOT
   2 -   8 -171.7   70.6  -98.6  -90.4 -163.6  2.027   1.62   0.27   0.06   1.
   3 -  16  -48.5  -51.9  -83.1  -62.6  -64.9  2.029   0.47   0.11   0.20   0.



TOTAL NUMBER OF DISULFIDE BONDS FOUND IN THIS PROTEIN     2




COMPARISON BETWEEN OBSERVED AND CALCULATED C-BETA POSITIONS

AVERAGE DISCREPANCY BETWEEN CBC AND CBO IS :   0.091
STANDARD DEVIATION OF DISCREPANCY IS :         0.065
LARGEST DEVIATION IS :                         0.238 IN RESIDUE    3


THE FOLLOWING RESIDUE PAIRS MIGHT FORM A DISULFIDE
BRIDGE ACCORDING TO THEIR CB-CB DISTANCE. WHICH
LIES  BETWEEN 3.27 AND 4.38 ANGSTROM
CORRESPONDING TO CHI-3 ANGLES OF +/- 90, +/- 47.2 DEGREES



  NR  RES1 --  RES2  NAME1 NAME2   CB DIST  CA DIST

   1    2  --    8    CYS   CYS     3.960    5.012
   2    3  --   16    CYS   CYS     4.060    5.398
   3    5  --    8    LEU   CYS     4.166    5.483

   3 PAIRS FOUND

AFTER CHECKING FOR CB-SG AND SG-SG DISTANCES THE
FOLLOWING RESIDUE PAIRS ARE STILL POSSIBLE


       SGDIST   X1     X2     X3    X2'    X1'  CHINRG TAUNRG DISNRG TOTNRG



  1 CONFORMATION BETWEEN :    2 -    8    CYS CYS

    1  2.030 -164.7   56.5  -96.9  -88.2 -157.5   2.05   0.03   0.04   2.13
    2  2.030 -154.3   36.9  -86.2  -93.8 -148.2   3.95   0.89   0.09   4.93
    3  2.031 -178.9    1.3  105.4   98.2  106.9   5.44   1.58   0.02   7.04


  2 CONFORMATION BETWEEN :    3 -   16    CYS CYS

    1  2.030  -66.2  -30.3 -103.6  -51.8  -77.9   1.84   0.15   0.02   2.01
    2  2.030 -155.3   29.9   80.6   55.5  175.0   2.00   1.23   0.13   3.36


  3 CONFORMATION BETWEEN :    5 -    8    LEU CYS

    1  2.030   33.5   56.6  104.6 -146.8   68.2   2.60   0.10   0.02   2.73
    2  2.030  147.1 -142.5   76.7   79.5  -46.0   3.70   0.10   0.23   4.03
```

Figura 33 - Trecho do resultado da execução do SSBOND

É importante ressaltar que os algoritmos descritos anteriormente propõem as possíveis pontes através de cálculos matemáticos considerando ângulos, posição da cisteína, quantidade de resíduos entre os pares, dentre outros. Nenhum deles considera a conformação total da cadeia principal como aqui proposto, ou seja as coordenadas de cada um dos átomos que compõe a cadeia principal para sobreposição com os registrados no banco.

5.1.4 Etapa 4 – Otimização da base de dados e da busca

Algumas decisões, otimizações, mudanças e aperfeiçoamentos foram feitos na base de dados inicial com o objetivo de tornar o sistema mais eficiente computacionalmente ou aumentar e até mesmo melhorar o banco de dados. Estas alterações e decisões seguem descritas nos itens seguintes:

5.1.4.1 Geração de arquivo para mais de uma cadeia

Cada arquivo PDB representa uma estrutura e este arquivo pode possuir várias cadeias da mesma estrutura que repetem a mesma interação, com pequenas variações na conformação, o que pode estimular a redução da quantidade de arquivos de interações pela eliminação dos arquivos dessas interações redundantes provenientes de uma mesma estrutura, com variação da cadeia. Porém, observando-se a Tabela 8, pode-se concluir que a manutenção dessas outras cadeias se faz necessária uma vez que a variação das conformações, apesar de ser pequena, pode resultar em sobreposições melhores do que outras cadeias da mesma estrutura.

5.1.4.2 Verificação de 16 distâncias

Para a redução da quantidade de pares encontrados na varredura da proteína alvo, a verificação das 16 distâncias entre os átomos N, CA, C e O (conforme indicadas na Figura 34), permite o aumento da quantidade de parâmetros a serem usados para a busca dos pares de resíduos candidatos. Inicialmente foram testadas apenas as distâncias entre átomos pares num total de 4 distâncias (N-N, CA-CA, C-C, O-O).

A variação das conformações das cadeias dos resíduos interagentes é muito grande, podendo se apresentar paralelos diretos, paralelos invertidos, perpendiculares, diagonais, dentre outras formas. Com isso, a variação dos valores das distâncias entre os átomos tende a uma faixa muito grande fazendo com que todas as distâncias possíveis aumentem os limites. Neste caso,

o uso das 16 distâncias permite refinar a busca eliminando pares candidatos que não satisfaçam a todas as faixas de valores para estas distâncias.

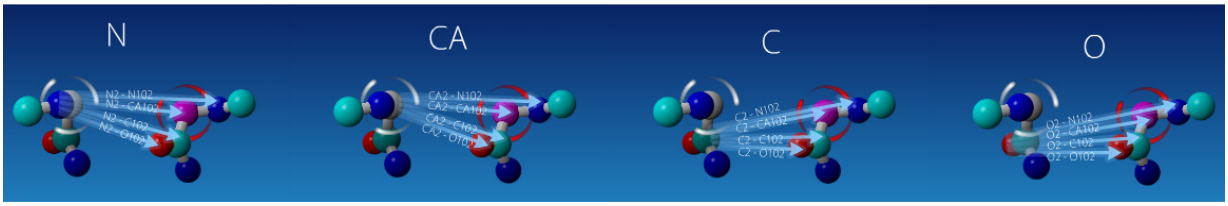


Figura 34 - 16 distâncias entre os átomos da cadeia principal dos resíduos centrais

5.1.4.3 Critério para sobreposição

Houve uma preocupação inicial na definição do critério para decidir quando uma sobreposição pode ser considerada boa o suficiente a ponto de recomendar a mutação, ou seja, o quão afastados do zero os valores da sobreposição ainda são válidos para se considerar a mutação. Com relação a esta métrica, chegou-se à conclusão de que o pesquisador, usuário da ferramenta, é que considerará a mutação, visualizando a sobreposição da interação candidata do banco sobre a proteína-alvo, como apresentado na Figura 36 e na Figura 37. Isto porque somente o usuário poderá definir se aquela interação pode ou não atender aos seus propósitos, devido aos átomos e/ou resíduos participantes, dentre outras características.

5.1.4.4 Integração, arquivos, código

A integração de algumas etapas ou módulos desenvolvidos visou a redução do número de interações com o banco ou com a lista de arquivos gerada, reduzindo o tempo de processamento. Arquivos intermediários também foram eliminados com esta integração e os dados gravados neles foram transferidos para o banco, para melhor gerenciamento. O algoritmo em alto nível foi apresentado no capítulo Metodologia.

Por outro lado a redução da quantidade de arquivos de saída da etapa de sobreposição, eliminando os arquivos “.log”, “.rmstab” e “.out”, utilizados para a validação da sobreposição e montagem da base de dados inicial também foi um fator que permitiu a diminuição do tempo de processamento e geração dos arquivos das interações. Isto resultou em uma redução drástica na quantidade de gravações no disco e do espaço ocupado pelos arquivos gerados. O

total de 39,4 GB ocupados por todos os 7.957.772 arquivos, inicialmente, foi reduzido para 7,8 GB ocupados pelos 1.989.443 arquivos ".deltas".

O uso da programação paralela para execução das sobreposições e geração da base de dados inicial foi testada no Cluster Veredas, do CENAPAD/UFMG, quando os arquivos das novas interações inseridas no banco foram gerados. O uso do cluster permitiu a execução simultânea de vários processos, sendo cada um deles para a geração de todos arquivos de uma interação.

A partir de um acesso ao cluster (via conexão ssh) um processo é submetido ao escalonador, que é o gerenciador dos processos em execução. O Veredas utiliza o slurm, um escalonador de processos que controla as filas de nós (*cores*) disponíveis para uso e seu hardware. Assim o escalonador pode enviar o processo submetido à execução a uma fila específica e consequentemente a um nó. As filas são organizadas por tempo máximo de processamento e a quantidade de nós de cada fila varia conforme a definição da administração do cluster, no caso do Veredas as filas são *short* (1 hora, 10 nós), *long* (4 dias, 72 nós), *superlong* (10 dias, 24 nós). O processo submetido, agora alocado a um dos nós, pode submeter novos *jobs* (outros processos) na fila de processos e estes serão alocados quando houver disponibilidade nas filas solicitadas.

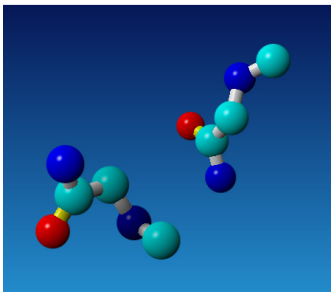
O código escrito para o Veredas consistiu em um processo principal submetido à fila *superlong* e este por sua vez inicializava os processos secundários para a geração dos arquivos de interação (fila *short*) ou sobreposição dos arquivos gerados (*long*). Cada processo desses inicia outros processos a partir do seu código, que geram os arquivos individualmente ou fazem a sobreposição desses arquivos gerados com os arquivos do banco ou inserem dados obtidos no banco de dados. Ao ser executado num *core* o processo executa os outros processos relacionados com o seu código em paralelo rodando localmente, acessando memória e discos locais do *core*. Os arquivos gerados são transferidos posteriormente para um disco compartilhado, mais especificamente para a pasta de trabalho do usuário do cluster, evitando o armazenamento no nó, que se torna inacessível após a finalização do processo.

Infelizmente os trabalhos desenvolvidos no cluster Veredas tiveram que ser interrompidos devido à problemas físicos relacionados com o resfriamento e a estrutura comprometida de alguns nós e do disco compartilhado. Esta interrupção atrasou a finalização do projeto, mas não o impediu, houve a necessidade de retornar o processamento ao servidor inicial.

5.1.4.5 Otimização da busca

A busca na base de dados gerada se tornou computacionalmente cara, ou seja, o número de sobreposições a serem realizadas para a verificação de cada par implica no aumento do tempo total da busca. Dessa forma, a próxima etapa do projeto consistiu em estudar uma forma de reduzir o tempo gasto e agilizar as buscas.

Os arquivos que compõem o banco (como exemplo a visualização na Figura 35) possuem coordenadas atômicas que podem conter pequenas variações em alguns átomos apenas, distinguindo umas das outras.



```
CRYST1   40.840   18.470   22.340  90.00  90.80  90.00 P 1 21 1      2
SCALE1      0.024486  0.000000  0.000342        0.00000
SCALE2      0.000000  0.054142  0.000000        0.00000
SCALE3      0.000000  0.000000  0.044767        0.00000
ATOM      1  C   CYS A   1      15.001   8.824  10.393  1.00  2.77     C
ATOM      2  N   CYS A   2      14.095   9.609   9.803  1.00  2.50     N
ATOM      3  CA  CYS A   2      14.156  11.056   9.974  1.00  2.29     C
ATOM      4  C   CYS A   2      12.854  11.599  10.522  1.00  2.21     C
ATOM      5  O   CYS A   2      11.788  11.089  10.230  1.00  2.80     O
ATOM      6  N   PRO A   3      12.946  12.670  11.330  1.00  2.18     N
ATOM      7  C   GLY A 101      17.532  13.749  13.433  1.00  2.77     C
ATOM      8  N   CYS A 102      18.016  13.261  12.264  1.00  2.61     N
ATOM      9  CA  CYS A 102      17.753  11.884  11.890  1.00  2.45     C
ATOM     10  C   CYS A 102      18.486  10.949  12.876  1.00  2.71     C
ATOM     11  O   CYS A 102      19.338  11.384  13.654  1.00  3.20     O
ATOM     12  N   ILE A 103      18.143   9.671  12.777  1.00  2.71     N
END
```

Figura 35 - Exemplo de par que compõe o banco de dados para sobreposição (à direita o arquivo PDB gerado)

Isto significaria que uma clusterização e remoção de elementos similares, dependendo do parâmetro a ser utilizado poderia impedir que aqueles átomos fossem validados numa busca. Sendo assim, uma sobreposição de todos os elementos contra todos poderia indicar os conjuntos com dados muito próximos (como na Figura 36, por exemplo), formando grupos ou clusteres de elementos similares que poderiam ser excluídos diminuindo-se a base de possíveis pares a serem pesquisados e indicados como possíveis mutantes. Na Figura 36 pode-se perceber que os três arquivos sobrepostos possuem pontos em comum (átomos perfeitamente sobrepostos) e pontos com uma variação maior em algumas coordenadas distintas. O mesmo pode ser observado na Figura 29.

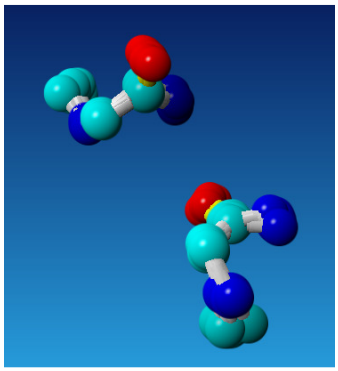


Figura 36 - Sobreposição de três arquivos PDBs muito similares do banco de dados gerados pelo algorítimo

Uma das otimizações consistiu no cálculo de um *score* a partir dos arquivos de deltas (Equação 6, descrita na Metodologia). A ideia é eliminar o tempo gasto com a sobreposição de uma estrutura com todas as outras que compõe o banco da interação, o que se repete para cada par candidato da proteína-alvo. Por exemplo, se o banco de interações possui 50.000 arquivos e a proteína-alvo possui 100 pares candidatos, serão 5.000.000 de sobreposições a serem executadas, caindo para apenas 100 com o uso do *score*. Sendo assim, os pares candidatos da proteína-alvo deverão ser sobrepostos apenas com um arquivo de referência, qual seja o arquivo de maior resolução usado na sobreposição dos pares encontrados na interação. Isto significa que pares candidatos e pares da interação serão comparados com base em uma mesma referência espacial e serão avaliados através do valor gerado pelo *score* de cada um.

A busca, portanto, será baseada em comparação numérica ao invés de cálculos de diferenças de sobreposição por átomo. Na busca usando sobreposição atômica, foram gastos pouco mais de 24 horas para o processamento do polipeptídeo 1pen (16 resíduos), conforme abordado anteriormente. Para a busca neste mesmo polipeptídeo usando o *score* foram gastos pouco menos de 1 minuto, pois o processamento necessário foi de apenas as sobreposições dos arquivos de candidatos gerados para o polipeptídeo contra a referência definida para a interação. O cálculo do *score* envolve a contribuição de cada distância individual que compõe o arquivo de deltas com o objetivo de se ter um valor resultante que seja sensível à variação dos valores das distâncias apresentadas no arquivo referido. Com isso é possível avaliar uma pequena diferença no valor de uma distância específica, podendo-se perceber então sutis diferenças de conformação.

O uso desse *score* permitiu ordenar os arquivos a partir de cada sobreposição com o arquivo de maior resolução, exemplo na Figura 37. Nesta figura, observam-se os dois pares a serem sobrepostos (à esquerda e ao centro a figura). À direita da figura encontram-se os pares sobrepostos. Nesta sobreposição é possível perceber que o valor do *score* será alto uma vez que há parcelas de distâncias significativas resultantes das diferenças das conformações dos átomos (indicadas como os tracejados à direita da figura). O conjunto de *scores* gera um vetor a ser utilizado numa busca binária[41] a partir de um par encontrado na proteína alvo (o resumo das etapas da busca se encontra na Tabela 10).

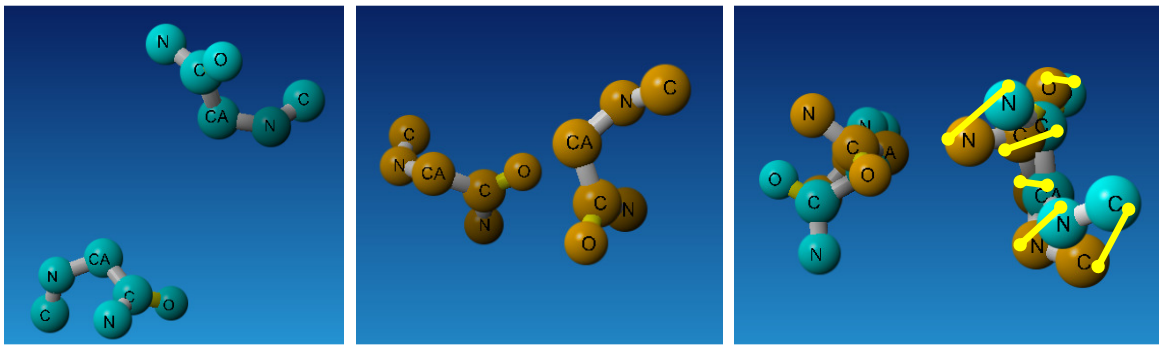


Figura 37 - Exemplo de sobreposição (direita) com indicação das distâncias; dois pares sobrepostos (esquerda e centro)

A seguir, um resumo das etapas de geração do banco de dados e da busca em uma proteína alvo. Estas etapas, detalhadas nas seções anteriores são apresentadas aqui de forma sucinta permitindo o acompanhamento das etapas.

A Tabela 9 apresenta o resumo das etapas da geração do banco de dados das interações atômicas. Este processo consistiu em levantar os dados dos arquivos obtidos no PDB para posteriormente gerar novos arquivos com as interações pesquisadas segundo o padrão definido. Estes arquivos são sobrepostos com um arquivo escolhido (maior resolução) e posteriormente são gerados os *scores* que identificam cada arquivo e que serão usados para a busca de pares da proteína alvo.

[41] Definido por Knuth (1997) como um algoritmo de busca numa tabela ordenada iniciando pelo meio do intervalo e desconsiderando os intervalos em que o valor buscado não se encontra repetindo-se iterativamente até que o valor seja encontrado.

Tabela 9 – Resumo das etapas de geração do banco de dados

	Etapa	Descrição
1	Download dos arquivos do PDB	Download de todas as proteínas do PDB, resolvidas por Difração por Raios-X, atualizações momentaneamente manuais para não afetar os resultados encontrados.
2	Envio para o banco, dos dados dos arquivos PDB	Uma vez disponíveis, os arquivos são identificados e cadastrados na tabela de proteínas do banco de dados para que sejam usados nos cálculos e varreduras.
3	Arquivos avaliados em busca da interação	Para cada interação contida no banco, o arquivo é percorrido considerando os átomos que compõem a interação e suas características.
4	Arquivos gerados com a interação encontrada	Para cada par de resíduos que compõe a interação, e que são encontrados na proteína, é gerado um arquivo contendo estes resíduos e um resíduo anterior e um posterior a cada um, caracterizando a interação, como apresentado na Figura 20 em sua forma completa e na Figura 35 sua forma resumida. Cada interação é inserida na respectiva tabela do banco de dados
5	Sobreposição dos arquivos gerados contra o de maior resolução do grupo (referência)	A partir da geração de todos os arquivos, cada um deles é sobreposto com o de maior resolução, gerando um arquivo com as diferenças da sobreposição (deltas, Figura 28 e Figura 37).
6	Geração do vetor de distâncias dos deltas	Para cada arquivo delta é gerado um *score* (calculado a partir da distância euclidiana,Equação 6) para identificá-lo perante os outros e servir de métrica para busca a ser realizada, evitando-se a sobreposição um a um no momento da busca. Este *score*, juntamente com as características da interação são inseridos na tabela de pares de interações do banco de dados.
7	Levantamento das distâncias atômicas características da interação	Ao término da geração dos arquivos, são levantadas as 16 distâncias que caracterizam as interações, como abordadas no Gráfico 6 e no Gráfico 7, estas distâncias alimentam a tabela de interações no banco.

A otimização das etapas da geração do banco também resultou em mudanças nas etapas da busca numa proteína alvo, conforme descrito na Tabela 10. A partir das distâncias obtidas nos arquivos das interações gerados para o banco, pares candidatos da proteína alvo são identificados e novos arquivos gerados para, posteriormente, serem sobrepostos com o mesmo arquivo referência usado na sobreposição dos arquivos da interação. Após isso, o *score* do arquivo é calculado e usado na busca do melhor par da interação.

Tabela 10 – Resumo das etapas da busca numa proteína alvo

	Etapa	Descrição
1	Identificação de par na proteína alvo a partir das distâncias da interação	A partir das distâncias calculadas átomo a átomo, que identificam uma interação (item 7 da Tabela 9), todos os pares de resíduo de uma proteína alvo são verificados se estão de acordo com estas distâncias e, se encontrados, são gerados arquivos no formato PDB para cada par encontrado.
2	Sobreposição com a referência da interação (de maior resolução)	Cada arquivo gerado é sobreposto com o arquivo de maior resolução da interação contido no banco, gerando um arquivo de deltas para cada sobreposição.
3	Cálculo do valor do *score* de cada arquivo	A partir dos arquivos de deltas é calculado o score (Equação 6) de cada arquivo, que corresponde a um par de resíduos, para pesquisa no vetor. Este score, bem como características do par encontrado são inseridos na tabela par de proteínas, do banco de dados.
4	Busca binária no vetor de distâncias	O *score* calculado na etapa anterior é usado para uma busca binária no vetor de valores calculados na etapa 6 da geração do banco de dados (Tabela 9). Os valores maior e menor são inseridos na tabela de par de interação de proteína, no banco.
5	Exibição na interface web dos resultados encontrados	A partir da interface web do sistema os itens identificados na etapa anterior são exibidos através de gráficos de distâncias e exibição da estrutura 3D, para que o usuário possa avaliar a adoção da mutação do par proposta.

5.1.4.6 Diagrama Entidade Relacionamento

O Diagrama Entidade Relacionamento (DER) é utilizado para a definição das entidades que compõem o banco de dados e dos relacionamentos/associações que ocorrem entre elas (Elmasri e Navathe, 2000). No diagrama é possível indicar os tipos dos campos bem como os índices de cada tabela. Na Figura 38 pode-se observar o DER construído para o banco, remodelando as tabelas previamente criadas.

As principais tabelas do banco são:

- ✓ ***uuser***, que contém o cadastro do usuário que terá acesso ao sistema, necessária para o desenvolvimento seguinte que é do sistema web e para controlar o acesso ao sistema bem como gerar estatísticas de uso do mesmo.

- ✓ ***interactions*****,** que contém os dados da interação inserida, ou seja, a sequência de átomos, a distância a ser buscada (*cutoff*), o arquivo de maior resolução usado para a

sobreposição, o tipo da interação, além das 16 distâncias entre os átomos da cadeia principal.

- ✓ ***protein***, que contém os dados da proteína inserida para ser pesquisada, sendo: código pdb, título, número de pares candidatos e número de possíveis localizações para interações encontradas, além da data de inclusão no banco e o nome do arquivo.

- ✓ ***pprocess*****,** que também é necessária para o acesso via interface web, armazenando as interações que foram solicitadas em pesquisas ao banco, mantendo uma lista de processos a serem executados para se ter controle da ordem de execução e de quais foram executados.

Os relacionamentos dessas tabelas com as outras geram novas tabelas para, por exemplo, caracterizar os pares da proteína (***protein_pair***), os pares da interação (***interaction_pair***), as proteínas do usuário (***user_protein***), os pares sobrepostos proteína/interação (***protein_interaction***), dentre outras informações características.

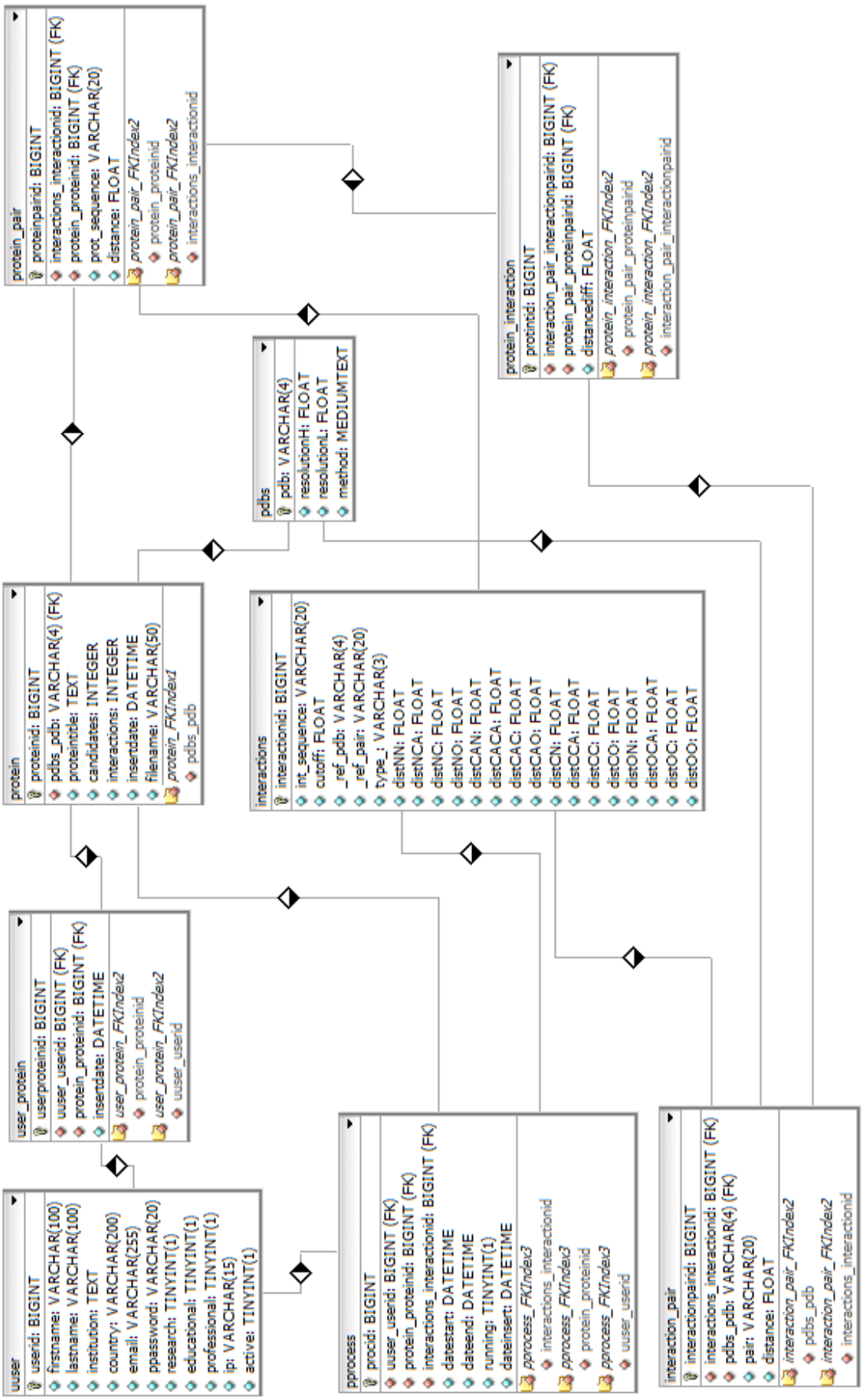


Figura 38 - DER do sistema

5.1.5 Etapa 5 – Inserção de novas interações no processo

Esta etapa, considerando as alterações e otimizações anteriores, teve como objetivo inserir novas interações a serem pesquisadas e propostas pelo algoritmo, como interações eletrostáticas e de ligações de hidrogênio. O princípio será o mesmo, ou seja, considerando-se a conformação das cadeias principais dos resíduos interagentes. Dessa forma, todo o algoritmo é mantido.

Após pesquisá-los no pdb para gerar o banco, as próximas etapas são gerar novos parâmetros de busca e identificação na proteína alvo, e por fim, a pesquisa na proteína alvo e a disponibilização dos resultados. Todas estas etapas, já implementadas para as pontes dissulfeto foram adaptadas para que se busque as novas interações (uma interação entre serina e serina, serina e treonina, aspartato e glutamato, aspartato e lisina, dentre outras). Ao todo foram inseridas mais 180 interações além da ponte dissulfeto, descritas na Tabela 4, anterior.

5.1.6 Problemas encontrados em arquivos do PDB

Nas várias etapas do projeto vários erros nos arquivos PDB foram encontrados o que ocasionou erros nos resultados obtidos, forçando a alteração dos códigos a fim de adaptar-se aos problemas encontrados. A Tabela 11 apresenta uma lista com os problemas encontrados, a identificação dos arquivos envolvidos, a descrição do erro, sua consequência e a decisão tomada para contornar a situação.

Como um exemplo disso, as distâncias 0 Å, 34,96 Å e 98,02 Å foram informadas em alguns arquivos PDB como distâncias que caracterizam a ponte registrada através do rótulo “SSBOND”. Isto denota um erro na anotação desses arquivos por estarem muito fora da faixa (34,96 Å e 98,02 Å são valores extremamente grandes para a distância entre dois enxofres que interagem).

Tabela 11 - Erros encontrados nos arquivos do PDB

Local	Alguns exemplos de arquivos envolvidos	Linha(s) de exemplo extraída(s) de arquivo(s) PDB		
		Descrição do erro	**Consequência**	**Decisão tomada**
Resolution high	-	Valor de Resolution High trocado com Resolution Low.	Arquivo não foi computado por não ter resolução suficiente.	Arquivo ignorado.
Número do resíduo	1a5i, 1bda	`ATOM 11 HG1 THR A 1A 60.378 20.455 43.183 0.00 0.00 H` `ATOM 12 N CYS A 1 59.576 25.193 41.879 1.00 43.33 N`		
		Alguns resíduos consecutivos foram encontrados numerados com o mesmo valor.	Na busca pelo próximo resíduo, a geração do arquivo foi comprometida pois o próximo resíduo estava muito distante do resíduo central.	Para a geração dos arquivos foi considerada a distância entre os resíduos, ou seja, validada a ligação peptídica.
ChainID	1afa, 1afb, 1afd, 1bch, 1gxv, 1ilr, 1k2i, 1kmb, 1kza, 1rtm, 1tmf, 1v9u	`SSBOND 4 CYS 2 191 CYS 2 220 1555 1555 2.05`		
		O identificador da cadeia (ChainID) foi anotado numericamente ao invés de letras.	Na geração dos arquivos não se distinguia a cadeia do número do resíduo, pois os mesmos eram consecutivos e não separados por marcador.	Arquivo ignorado.
Nome do resíduo	1czq	`SSBOND 1 DCY D 3 DCY D 14 1555 1555 2.05`		
		Nomes diferentes para os resíduos, possivelmente por dupla conformação ou outra anotação diferenciada.	Resíduos padrões (aminoácidos essenciais) não foram identificados prejudicando a composição do banco com a ponte existente no arquivo, porém anotada incorretamente.	Arquivo ignorado.
Altloc	1c4y, 1czq, 1deu	`SSBOND 1 CYS A 10P CYS A 31 1555 1555 2.53`		
		Número que identifica o resíduo anotado juntamente com uma letra, em alguns casos, vários resíduos consecutivos com a mesma sequencia.	Erro na identificação dos resíduos consecutivos para geração do arquivo contendo seis resíduos.	Arquivo ignorado.

SSBOND	2opw, 1rbo, 1nty, 2k6d	`SSBOND 1 CYS A 3 CYS A 3 1555 4555 98.02`
		`SSBOND 3 CYS E 247 CYS E 247 1555 3555 39.46`
		`SSBOND 1 CYS A 1313 CYS A 1313 1555 1555 0.00`		

		Valores improváveis para a distância entre os átomos de enxofre de uma ponte dissulfeto.	Se a distância foi calculada incorretamente, causará erros nas estatísticas e validação das pontes identificadas.	Arquivo ignorado.
SSBOND	2bvr	`SSBOND 1 CYS H 42 CYS H 58 1555 1555`		
		Ausência do valor da distância entre os átomos de enxofre de uma ponte dissulfeto.	A ausência da distância causará erros nas estatísticas e validação das pontes identificadas.	Arquivo ignorado.

5.2 Desenvolvimento do sistema

Como produtos desse trabalho temos a base de dados desenvolvida e o sistema para acesso à mesma via *web*. O objetivo desse sistema é permitir o acesso amplo à base de dados pela comunidade científica e portanto, está disponível a partir da seguinte URL:

http://www.bioest.icb.ufmg.br/RID

A interface do sistema é apresentada em inglês para possibilitar o uso do sistema por pesquisadores de outros países. O primeiro acesso ao sistema (Figura 39) remete à tela de *login* ou, caso não tenha sido realizado um cadastro prévio, é possível cadastrar um novo usuário no link *NewUser* na tela principal.

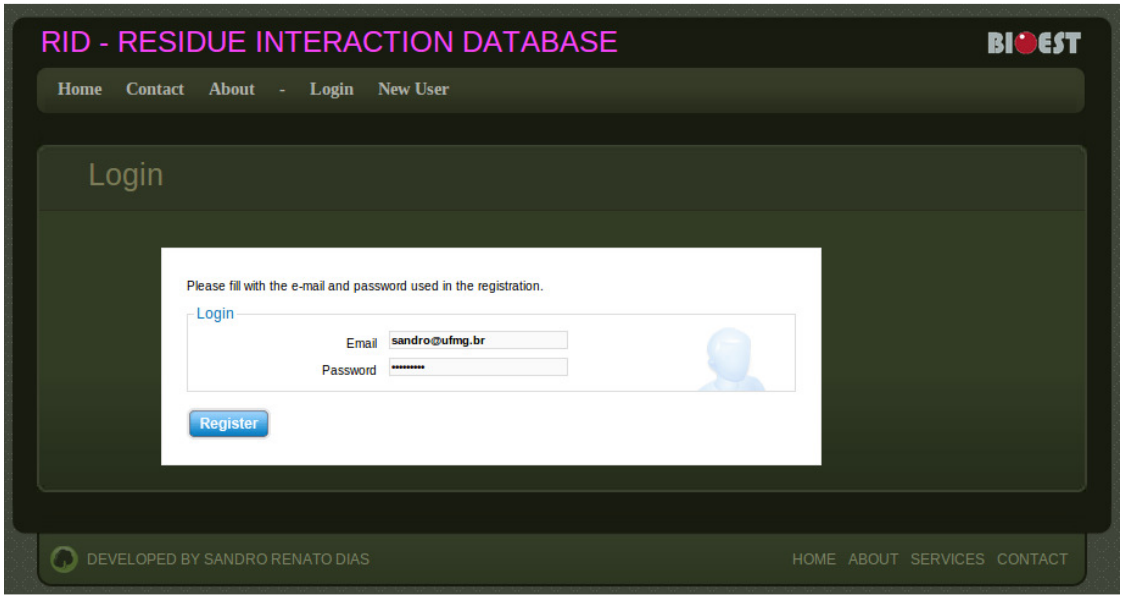


Figura 39 - Tela de *login*

O menu do sistema, antes da validação do usuário, possui as opções *Home* – para a tela inicial, *Contact* – contato com os responsáveis, *About* - para a descrição do projeto, além de *Login* e *NewUser*. Após a validação, são acrescentadas as opções *MyProteins* – para que o usuário visualize as proteínas alvo já submetidas por ele, *SubmitProtein* – para submissão de novos arquivos alvos, *Statistics* – para visualização de dados estatísticos do sistema, *Jobs* – para a visualização dos processos em execução, além de *Logout* – para sair da área do usuário. No caso de submissão de novos arquivos do usuário, ao invés de indicação de proteínas do

PDB, estes arquivos serão gravados no servidor e ficarão disponíveis apenas para o usuário que os submeteu, garantindo a confidencialidade dos seus dados e pesquisas. No caso de proteínas indicadas do PDB – banco público -, os seus resultados ficarão disponíveis para qualquer outro usuário que a indicar, ou seja, se um usuário resolver incluir a proteína X, já pesquisada por outro usuário, este novo usuário já a terá em sua lista com os resultados até então processados para a proteína.

5.2.1 Integração de tecnologias

Para o desenvolvimento do sistema, algumas tecnologias foram utilizadas e integradas com o objetivo de agregar funcionalidade e maior usabilidade ao desenvolvimento usando PHP, HTML, MySQL, Apache. Dentre estas, destacam-se o uso de:

I. A linguagem ***JavaScript*** foi usada para validação dos campos dos formulários, em termos de tamanhos, tipos e quantidades de itens selecionados além da manipulação dos gráficos gerados (vetores de dados a partir dos dados do banco) e dos *scripts* de visualização das moléculas (jmol *scripts*).

II. A ferramenta **jmol**[42], através do seu *applet* de código aberto, foi usada como ferramenta de visualização de estruturas online, permitindo uma personalização da visualização através de *scripts* dentro da própria página que a está exibindo, incorporando ao código da página o script da visualização possibilitando dinamismo como uso dos arquivos recém-processados em PHP para serem visualizados em três dimensões.

III. Biblioteca de gráficos ***HighCharts***[43] que possibilitou uma melhor interpretação das distâncias e dos candidatos a mutantes, exibidos através dos gráficos de colunas e dispersão. Os gráficos são gerados através de códigos em *javascript* que tiveram que ser adaptados para receber os valores provenientes do banco, gerando gráficos dinâmicos de acordo com as demandas do código e do banco, em função dos cliques

[42] Jmol: an open-source Java viewer for chemical structures in 3D. Disponível em: http://www.jmol.org.
[43] Disponível em: http://www.highcharts.com/

do usuário. O uso de gráficos de dispersão possibilitou avaliar o grau de similaridade entre a estrutura do par de aminoácidos mutante e a do par da proteína nativa.

IV. O uso de **CSS** com **HTML** permitiu não só a padronização da interface como a facilidade do desenvolvimento sem a preocupação de detalhes como cor ou layout de várias estruturas (tabelas, menus, botões, formulários, dentre outros).

V. O banco **MySQL**, alimentado pelos códigos *shellscript Bash* que inserem as interações é usado na interface web para acesso a estas interações, validação e registro do usuário, bem como a submissão das proteínas novas para serem analisadas.

VI. Além dos códigos em ***shellscript*** para geração dos arquivos das interações e montagem do banco, um código em *shellscript* roda permanentemente no servidor para acompanhar a execução dos processos de proteínas submetidas, organizando e gerando os resultados no banco. Ele funciona como um escalonador das execuções das sobreposições e geração dos arquivos da proteína nova ou de novas interações para uma proteína já existente.

VII. Uso da API **CAPTCHA**[44] do Google para validação do formulário foi necessária para evitar o uso ou o acesso indevido ao sistema por robôs e *scripts* automatizados na web. Esta validação ocorre apenas no registro do usuário. Esta API é disponível para uso mediante registro e download de alguns códigos e validação na base de dados do Google, pois um dos propósitos do seu desenvolvimento é não só colaborar para a segurança, mas também para a digitalização de documentos antigos com a verificação por humanos de palavras que não foram reconhecidas computacionalmente.

5.2.2 Uso do sistema

Para uso do sistema é necessário ter um usuário e senha, que permite a submissão de arquivos PDB ou o uso de arquivos já referenciados por outros usuários. O cadastro é simples (Figura 40) consistindo do primeiro nome, último nome, instituição de vínculo, país e intenção de uso

[44] CAPTCHA (Completely Automated Public Turing Test To Tell Computers and Humans Apart). Disponível em: http://www.google.com/recaptcha/captcha.

(pesquisa, educativo e profissional), e-mail, senha, além do captcha, necessário para evitar cadastros provenientes de *scripts* automáticos.

Figura 40 - Formulário de registro de usuário no sistema

Após o cadastro, o usuário poderá indicar um arquivo PDB através do seu ID (código PDB) ou submeter o próprio arquivo no formato PDB (seguindo o descrito na seção 1.2.1 e na seção 1.2.2) para avaliação da ferramenta (Figura 41). Caso escolha a primeira opção, após informar o código no campo apropriado, o arquivo correspondente àquele código será transferido do repositório do PDB para o servidor da aplicação através da URL http://www.pdb.org/pdb/files/$pdbcode.pdb, onde $pdbcode é o código informado pelo usuário no formulário. Após a transferência do arquivo o mesmo estará associado à conta do usuário, constando na tabela *proteins*, que mantém o cadastro das proteínas, e na tabela user_protein, que mantém o registro das proteínas usadas pelo usuário. Caso o arquivo já exista na base e já tenha sido avaliado anteriormente, o usuário terá acesso aos dados já disponíveis desse arquivo, que neste momento estará associado ao usuário anterior e a este novo usuário. Caso o código seja incorreto ou o arquivo não esteja mais disponível via URL, uma mensagem de erro será disponibilizada e o usuário deverá retornar ao formulário ou selecionar um arquivo da sua lista (Figura 42).

Caso a escolha do usuário seja a segunda opção, submeter um arquivo PDB, o mesmo poderá ser indicado através do campo apropriado do formulário e após sua transferência para o servidor, o mesmo estará disponível em sua lista de arquivos.

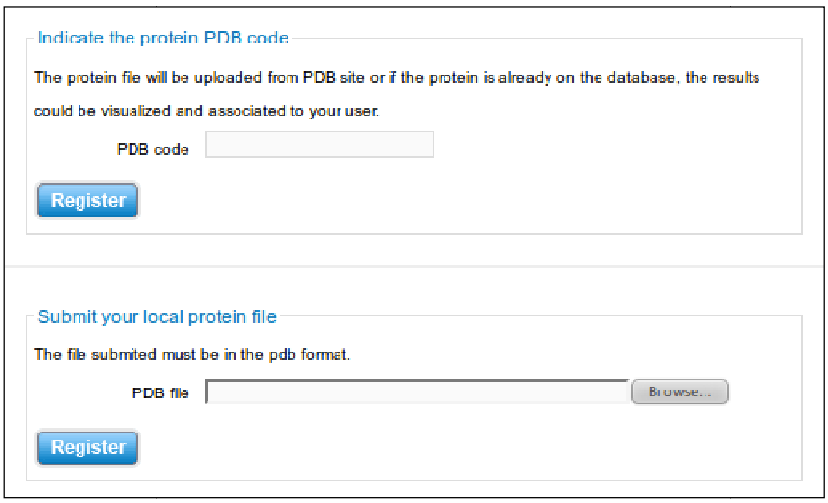


Figura 41 - Formulário de submissão de arquivo ou indicação do código PDB

5.2.3 Arquivos submetidos pelo usuário

A lista dos arquivos do usuário (Figura 42) apresenta uma tabela contendo os arquivos submetidos ou indicados pelo usuário, o título contido na identificação do arquivo, o número de candidatos já indicados e o número de interações finalizadas. Nesta tabela também é possível remover o arquivo da lista de arquivos do usuário, porém o mesmo continuará no sistema uma vez que pode ser procurado por outro usuário ou que já esteja na lista de outro usuário. Clicando-se no valor da coluna *Number of interactions finished* do arquivo é possível escolher as interações que serão pesquisadas na proteína (Figura 43). Já clicando-se no código PDB da proteína é possível visualizar as interações já pesquisadas (Gráfico 8).

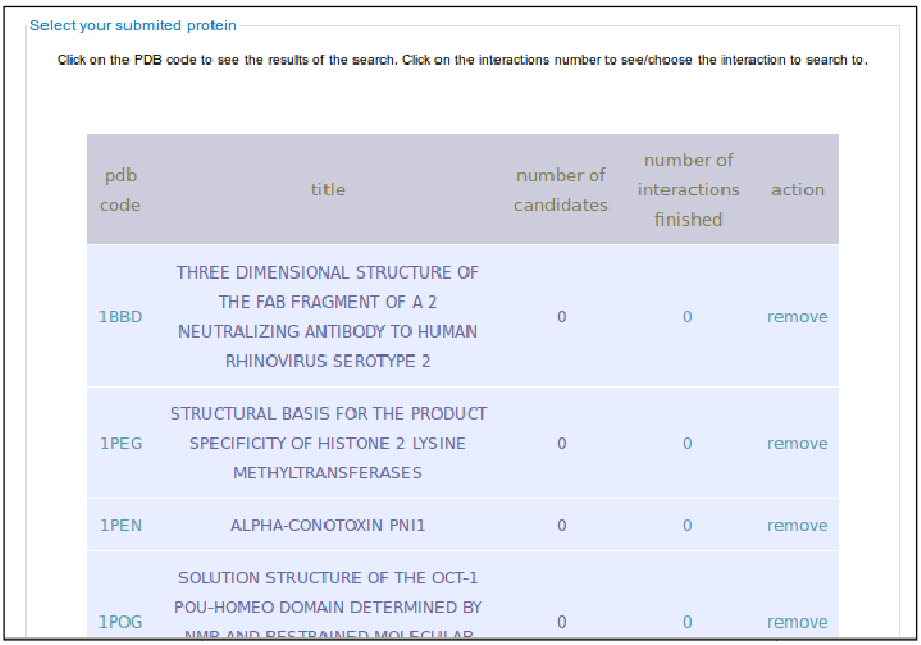


Figura 42 - Lista dos arquivos do usuário indicando a quantidade de interações já concluídas e o número de candidatos encontrados

5.2.4 Escolha de interações

A Figura 43 apresenta a relação das interações existentes no banco e que podem ser escolhidas pelo usuário para serem utilizadas no arquivo selecionado (o código do arquivo aparece no final do texto logo abaixo do menu, 1POG[45]). As caixas de marcação (*checkboxes*) de cada interação podem ter os seguintes estados: marcada e desabilitada, indicando que a interação já foi executada; marcada, indicando que a interação está sendo executada sobre o arquivo ou está aguardando para ser; desmarcada, indicando que a interação não foi escolhida pelo usuário. Ao se clicar em *Search* as interações escolhidas são inseridas no banco, para serem posteriormente processadas/pesquisadas na proteína, e retorna-se à tabela de proteínas do usuário (Figura 42).

[45] PDB 1POG. Cox, M., van Tilborg, P.J., de Laat, W., Boelens, R., van Leeuwen, H.C., van der Vliet, P.C., Kaptein, R. Solution structure of the Oct-1 POU homeodomain determined by NMR and restrained molecular dynamics. J.Biomol.NMR 6: 23-32. PubMed: 7663141. DOI:10.2210/pdb1pog/pdb, 1995.

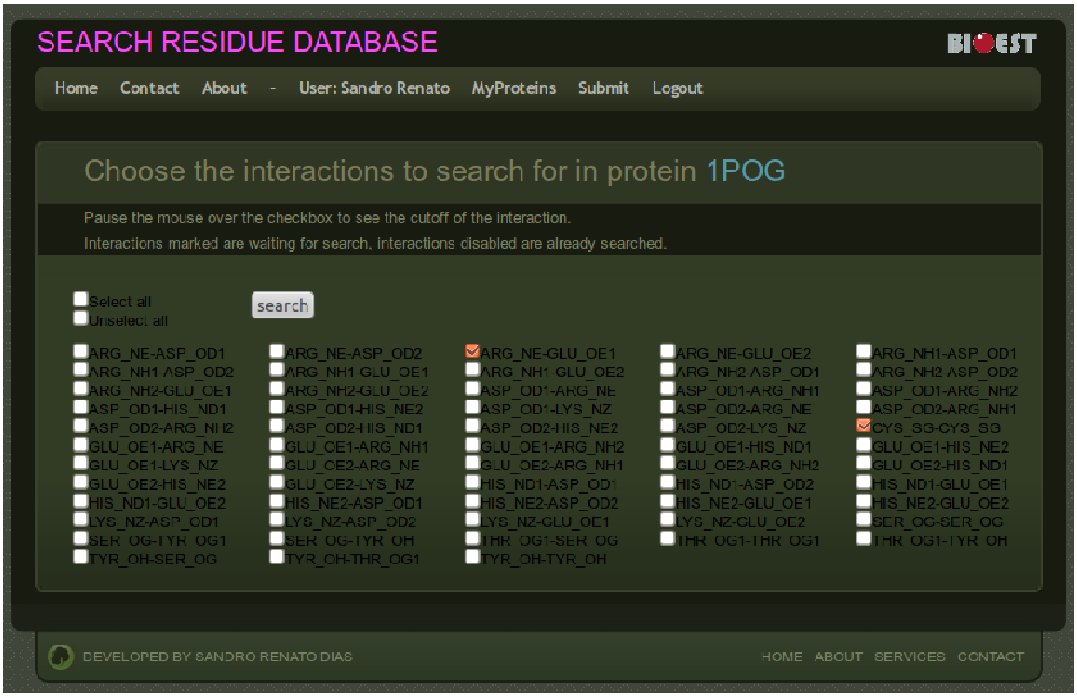


Figura 43 - Escolha das interações

5.2.5 Visualização das interações

O Gráfico 8 apresenta as interações já pesquisadas para a proteína, no caso foi escolhida apenas a ponte dissulfeto. Os pontos do gráfico representam os pares do banco que tem conformação muito próxima da conformação da proteína. No caso do gráfico em questão está representado apenas um par de interação do banco para cada par da proteína, sendo possível aumentar este valor a partir da interface.

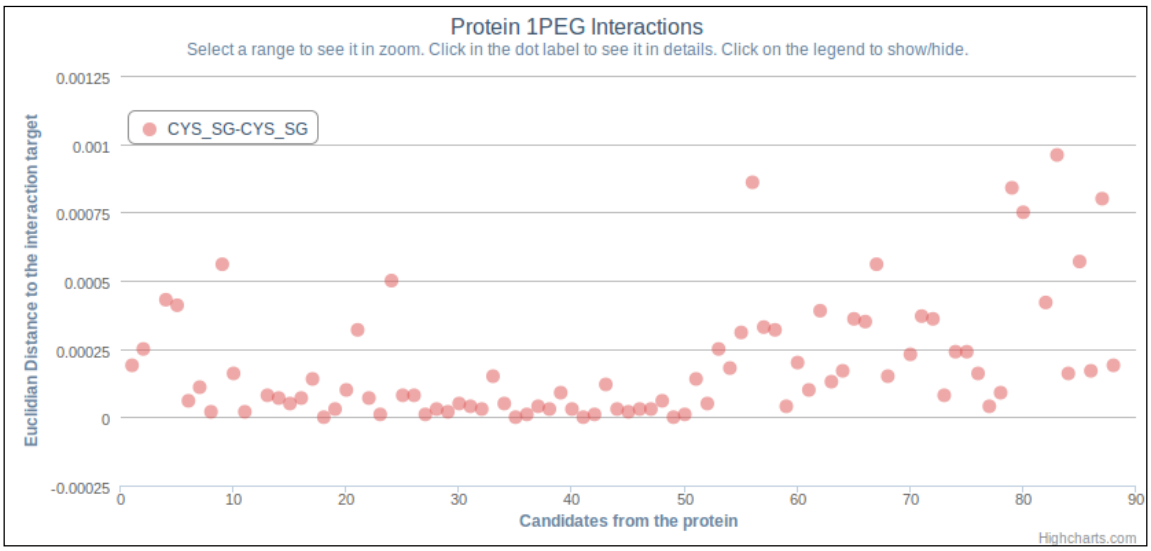


Gráfico 8 – Potenciais interações (pontes dissulfeto) a serem inseridas na proteína 1PEG após introdução da dupla mutação sugerida

Clicando-se na área do Gráfico 8 e selecionando-se uma região é possível visualizar com mais detalhes os pontos daquela região. O Gráfico 9 apresenta uma seleção realizada mostrando alguns pontos apenas. Um ponto em destaque, à extrema direita do gráfico, ao ser clicado, permite visualizar a identificação do par da proteína (59) e a diferença entre os *scores* (valor calculado a partir da Equação 6) calculados para a proteína e para a interação do banco. A diferença entre os *scores* em Angstrons (Å), portanto, é o eixo da ordenada do gráfico. Quanto menor esta diferença, mais próxima é a conformação da cadeia principal do par de aminoácidos da proteína alvo (submetida pelo usuário) com a conformação da cadeia principal do par de aminoácidos interagentes contido no banco de dados e sugerido pelo sistema. A vantagem do uso desse valor é que a busca a ser realizada no banco consiste em apenas uma consulta SQL[46] simples no banco hospedado no MySQL, sem a necessidade da sobreposição do par da proteína com todos os pares da interação para encontrar uma proximidade entre eles.

Ao se clicar na descrição que se abre quando o ponto é clicado é possível visualizar com detalhes a diferença entre a conformação da interação com o par da proteína, como observado no Gráfico 10.

[46] Structured Query Language – Linguagem de consulta utilizada em sistemas de gerenciamento de bancos de dados como o MySQL, por exemplo.

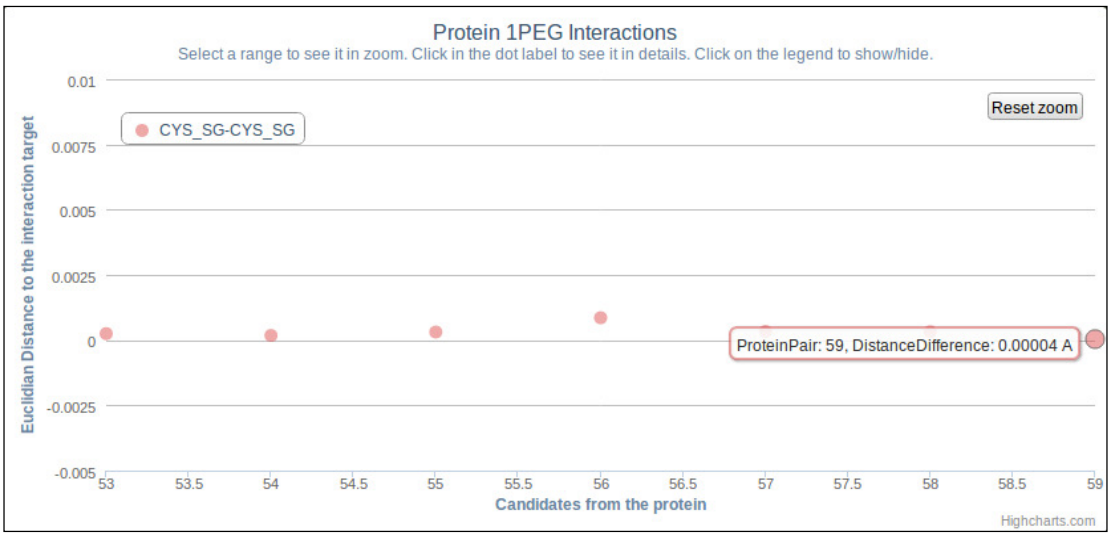


Gráfico 9 - Zoom em região da figura anterior, podendo-se observar com detalhes. Ponto clicado indicando os detalhes (identificação e valor).

O Gráfico 10 apresenta o ponto escolhido no Gráfico 9, anterior. Neste gráfico, o par da proteína é identificado como *Target*, o par da interação do banco é identificado como *Matched*; da mesma forma, os *Matched* combinados com -2, -1, +1 e +2 são dois elementos anteriores e dois elementos posteriores ao encontrado como melhor interação ou seja, menor diferença entre os valores de *score* para o par da proteína e o par da interação. A navegação entre os pares da interação, para serem observados com mais detalhes, pode ser feita a partir da Tabela 12.

O Gráfico 10 é composto de três modalidades diferentes e pode-se observar:

- ✓ No gráfico de pizza são plotados as porcentagens dos valores dos *scores* calculados para que sejam comparados em fatias, as dimensões dessas fatias podem servir de métrica para a comparação entre elas. Fatias azuis são do par da proteína e do par da interação, vermelhas referem-se aos pares imediatamente anterior e posterior e as fatias verdes pares a uma distância de dois elementos anteriores ou posteriores do par encontrado na interação.

- ✓ A curva plotada no gráfico apresenta os valores em Angstrons dos *scores* encontrados, visualizados sobre cada um dos elementos plotados no gráfico de barras. Para ocultar sua exibição basta clicar no nome respectivo (*Distance*) na legenda.

- ✓ O gráfico de barras apresenta, para cada barra a distância daquele átomo ao mesmo átomo do arquivo de referência, obtida após a sobreposição (dados provenientes do arquivo delta gerado). É possível ocultar uma ou mais barras neste gráfico bastando para isso clicar no respectivo átomo na legenda. Isto reorganizará o gráfico exibindo apenas as barras restantes.

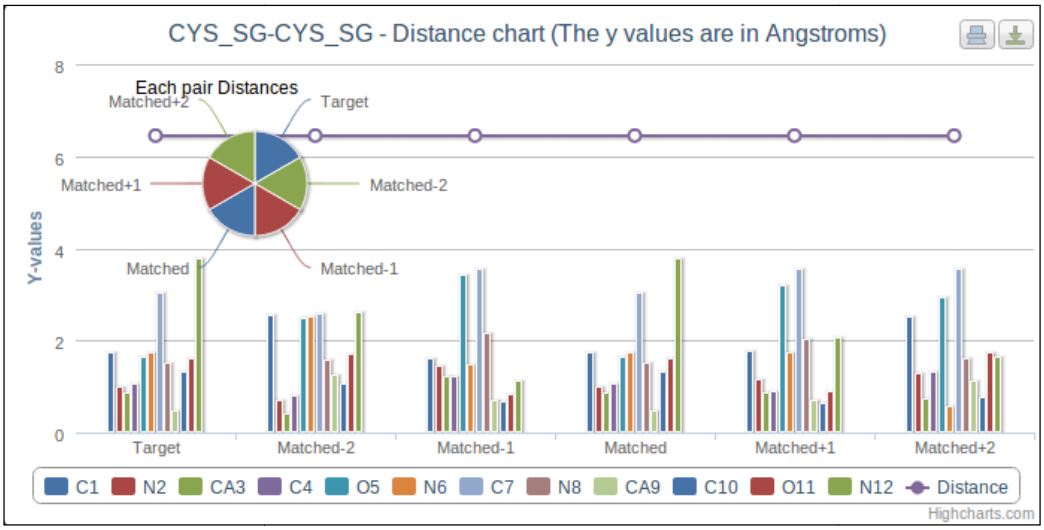


Gráfico 10 - Visualização do ponto clicado, podendo-se observar os detalhes das distâncias da interação encontrada (*Matched*) e do par da proteína-alvo (*Target*)

O Gráfico 11 apresenta um par diferente tanto para a proteína quanto para a interação, mudança esta que pode ser feita a partir da Tabela 12, onde se observa a diferença das alturas das barras além da curva e do gráfico de pizza.

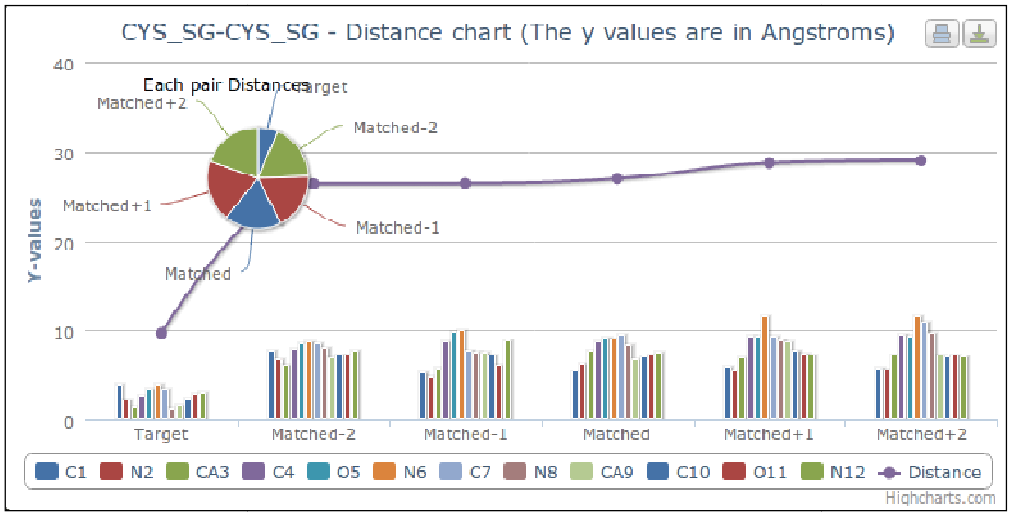


Gráfico 11 – Navegação no Gráfico 10 visualizando um par da interação que difere do par da proteína alvo, o que pode ser feito observando-se a sequência das barras

A Tabela 12 apresenta a lista dos pares da proteína-alvo e da interação que são listados no Gráfico 10. Ela é exibida na mesma página em que o gráfico é exibido, permitindo que o usuário possa navegar nos resultados e visualizar detalhes de cada um deles, como os resíduos que compõem a interação e o par da proteína alvo, o *score* calculado e a média dos arquivos delta, além da diferença do score calculada.

Na tabela também é possível visualizar a estrutura do par com o jmol, em outra janela, clicando-se em jmol *view*; também é possível visualizar o arquivo PDB referente à proteína daquele par, clicando-se em *pdb view*; já o link *superpose target* permite visualizar a sobreposição do arquivo daquele par selecionado com o *target* (par da proteína alvo), para tanto, a sobreposição, usando o software lsqkab do pacote CCP4 é executada e o arquivo gerado apresentado usando-se o jmol para a visualização da estrutura tridimensional. O link no código PDB da proteína permite abrir o site do PDB em outra janela para se visualizar os detalhes da proteína diretamente no site. O link sobre qualquer par da interação do banco permite posicionar aquele par como o *matched*, ou seja, o principal da interação para aquele par da proteína-alvo.

Tabela 12 - Busca do melhor par candidato à mutação

type	pdb	residue1	residue1 #	chain	residue2	residue2 #	deltas average	score distance	target distance diff	actions		
target	1per	CYS	2	A	TYR	15	2.6401	9.6159		jmol view	pdb view	superpose target
matched 2	1pe6	CYS	22	A	CYS	95	7.5876	26.4099	16.7640	jmol view	pdb view	superpose target
matched-1	1pem	CYS	167	A	CYS	178	7.4590	26.4566	16.8106	jmol view	pdb view	superpose target
matched	1pe6	CYS	25	A	CYS	95	7.7046	27.0038	17.3579	jmol view	pdb view	superpose target
matched+1	1peg	CYS	165	B	CYS	244	8.1317	28.7339	19.0880	jmol view	pdb view	superpose target
matched+2	1peg	CYS	165	B	CYS	308	8.1557	28.9901	19.3441	jmol view	pdb view	superpose target

A Figura 44 apresenta a visualização, em três dimensões usando o jmol, de três pares, da proteína-alvo (centro, *target*), da interação que se mostrou mais próxima ou que foi indicada

pelo usuário (*matched*, direita) e o par imediatamente anterior (*previous*, à esquerda). Esta figura também se apresenta na mesma página do Gráfico 10 e da Tabela 12 anteriores, e nela pode-se selecionar outra interação para ser visualizada para esta proteína.

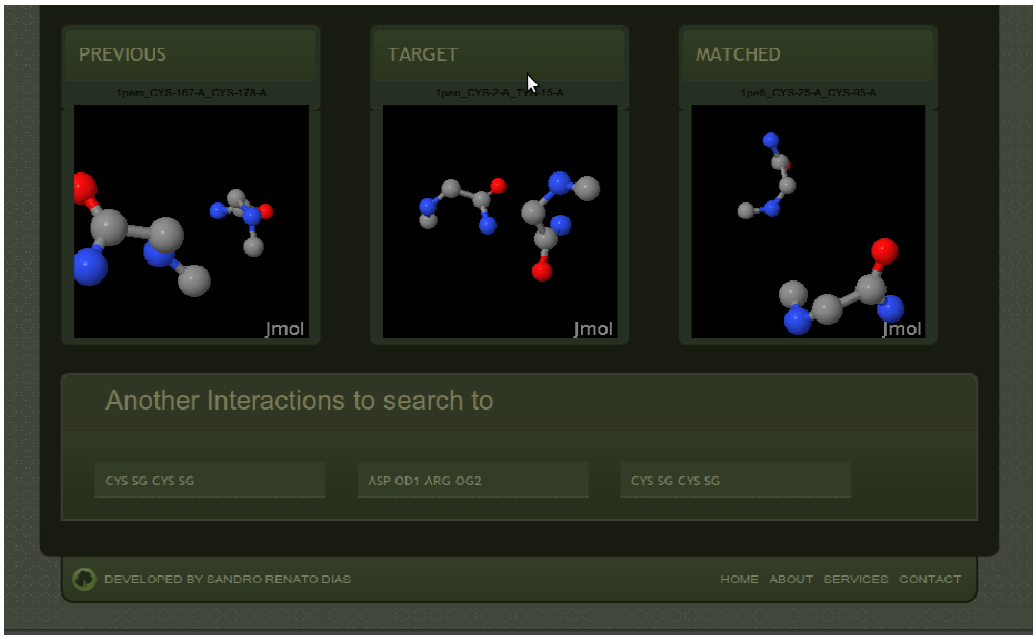


Figura 44 - Visualização dos três pares mais próximos

5.2.6 Análise da busca

Da mesma forma que foi feita uma busca para o polipeptídeo 1PEN na fase de desenvolvimento do banco de dados, a mesma busca foi realizada no desenvolvimento do sistema, porém o tempo de busca foi muito menor pois consistiu de uma consulta SQL simples ao invés de uma sobreposição de cada par candidato da proteína com todos os pares do banco. A sobreposição ocorreu, porém de todos os pares candidatos da proteína com o arquivo de maior resolução usado para a sobreposição dos arquivos da interação. Assim, para cada nova interação que for pesquisada, os pares candidatos deverão ser gerados novamente, pois dependem das distâncias que caracterizam a interação, e deverão ser sobrepostos com o arquivo escolhido como referência daquela interação. Este procedimento permitiu um ganho de desempenho grande no sistema, pois a busca agora ocorre de forma bastante simples, através de uma consulta SQL de poucos segundos.

A visualização gráfica dos pares também permite uma melhor identificação dos possíveis mutantes bem como a forma de comparação, baseada em uma equação (Equação 6) com contribuições de cada valor constante no arquivo deltas, diferente de usar-se apenas um dos parâmetros desse arquivo. Isto faz com que a avaliação do par candidato seja mais precisa considerando as pequenas variações de deltas, pois caso o maior valor do arquivo de deltas seja o mesmo para dois pares esta avaliação se torna mais complexa, para a busca anterior.

Com o objetivo de avaliar se as mutações propostas seriam interessantes e poderiam contribuir para aumentar a estabilidade da proteína, foram feitas algumas análises dessas mutações a partir de ferramentas online. Contudo, estas ferramentas consideram apenas as mutações individualmente, ou seja, a mutação de apenas um resíduo e não de dois resíduos concomitantes, com é proposto aqui. De qualquer forma, esta avaliação serve como um parâmetro para estimar a possibilidade de sucesso da mutação concomitante proposta.

A partir da análise de estabilidade da mutação para cisteínas do par ASN12A-TYR15A encontrou-se que esta mutação pode aumentar a estabilidade do polipeptídeo, conforme resultado da ferramenta MUpro[47] (Cheng, Saigo, Baldi, 2006) apresentada na Figura 45. MUpro é um conjunto de programas de aprendizado de máquina para predizer os efeitos da mutação de resíduos de aminoácidos na estabilidade da proteína.

[47] Disponível em: http://mupro.proteomics.ics.uci.edu

Mutation Request:
Name: 1penASN12
Sequence:

GCCSLPPCAANNPDYC

Position: 12
Original Amino Acid: N
Substitute Amino Acid: C

Prediction Results:

1. Predicted both value and sign of energy change using SVM and sequence information only (Recommended)

detal delta G = (INCREASE stability)

Mutation Request:
Name: 1penTYR15
Sequence:

GCCSLPPCAANNPDYC

Position: 15
Original Amino Acid: Y
Substitute Amino Acid: C

Prediction Results:

1. Predicted both value and sign of energy change using SVM and sequence information only (Recommended)

detal delta G = (INCREASE stability)

Figura 45 – MUpro - resultados encontrados na avaliação da estabilidade da mutação N12C e Y15C, do polipeptídeo 1PEN

Já a ferramenta AUTO-MUTE[48] (Masso, Vaisman, 2010), para a mesma mutação apresentou resultado similar, ou seja, o ganho de estabilidade com a mutação proposta, conforme relatório na Figura 46. AUTO-MUTE consiste de ferramentas web com o mesmo propósito que a ferramenta anterior, a predição de mudanças de estabilidade a partir de mutações simples em proteínas de estrutura nativa conhecida.

PDB_ID	Chain	Mutation	Stability	Confid	Vol	sT	Loc	Num	SS
1PEN	A	N12C	Increased	0.53	11.0	0.13	S	4	C
1PEN	A	Y15C	Increased	0.92	9.2	0.21	S	4	C

Figura 46 – AUTO-MUTE - resultados encontrados na avaliação da estabilidade da mutação N12C e Y15C, do polipeptídeo 1PEN

[48] Disponível em: http://proteins.gmu.edu/automute

A avaliação das mutações simples, indicadas pelas ferramentas anteriores é um indício de possibilidade e sucesso da mutação simples. Porém no caso de dupla mutação, o resultado pode variar uma vez que envolve dois novos resíduos que irão interagir entre si e com o restante no entorno. Isto pode ocasionar um sucesso para a mutação, com a adição da nova interação ou o fracasso com a interação de um dos resíduos mutados com outro resíduo diferente impedindo ou dificultando a interação inicialmente proposta. Estes resultados foram utilizados aqui apenas para uma avaliação rápida da possibilidade da mutação.

A ferramenta AUTO-MUTE também apresenta uma avaliação de mudança na atividade da proteína submetida através de seu formulário, a partir da mutação proposta. Os resultados, apresentados na Figura 47 apontam que as mutações propostas não afetam a atividade do polipeptídeo.

PDB_ID	Chain	Mutation	Prediction	Confid	Vol	sT	Loc	Num	SS
1PEN	A	N12C	unaffected	0.59	11.0	0.13	S	4	C
1PEN	A	Y15C	unaffected	0.73	9.2	0.21	S	4	C

Figura 47 – AUTO-MUTE - resultados encontrados na avaliação da mudança de atividade para a mutação N12C e Y15C, do polipeptídeo 1PEN

Estes resultados, baseados nas ferramentas MUpro e AUTO-MUTE, demonstram que o propósito do banco de dados de interações se confirma, pois a mutação avaliada, uma das mutações propostas por este trabalho para o polipeptídeo 1PEN, (substituição dos resíduos ASN12A e TYR15A ambos por CYS) possibilitam que a proteína aumente sua estabilidade conformacional sem afetar sua atividade, com a adição de uma nova ponte dissulfeto.

Outros pares propostos para o polipeptídeo 1PEN não tiveram a mesma avaliação pelas ferramentas, alguns pares tiveram os dois resíduos avaliados como contribuição negativa para a estabilidade e em outros pares apenas um dos dois. Como a avaliação das ferramentas é por mutação simples e não por uma mutação dupla simultânea estes resultados podem ser diferentes (para maior ou menor) de acordo com a mutação, uma vez que a mutação dupla visa a inserção de uma nova interação no polipeptídeo.

6. Considerações Finais

Os vários estudos sobre estabilidade de proteínas presentes na Literatura inspiraram o desenvolvimento desse trabalho, uma base de dados de conformações de pares de resíduos que interagem em proteínas de estrutura tridimensional conhecida. Com o objetivo de propor mutações em proteínas baseando-se em sua estrutura, a ferramenta apresentada neste trabalho, que consiste de um banco de dados de conformações de pares de resíduos, também apresenta um sistema, que pode ser acessado via web, para a análise de proteínas e verificação dos possíveis pares que possam ser mutados.

Como o objetivo da mutação a ser proposta pela ferramenta é manter a conformação da cadeia principal do resíduo mutado, alterando-se, portanto, apenas a cadeia lateral, isto pode resultar em um mutante "*in silico*" com possibilidades estereoquímicas de existir "*in vitro*". Esta ferramenta então pode ser usada para se inferir duas mutações concomitantes em uma proteína alvo, objetivando-se introduzir uma nova interação entre resíduos e aumentar a estabilidade conformacional e térmica da proteína mutante. Dessa forma, as mutações são propostas de forma a manter o enovelamento (e função), basicamente, conservando a conformação da cadeia principal dos resíduos mutados. Um ponto relevante a se destacar é com relação ao conteúdo do banco de dados, as conformações dos resíduos. Algumas ferramentas existem para propor mutações ou para identificar interações, algumas delas citadas neste trabalho, mas nenhuma encontrada utiliza a conformação da cadeia principal como parâmetro para a proposição de mutações. A base de dados aqui apresentada utiliza esta informação e justamente isto permite que a manutenção da estabilidade possa ocorrer.

As mutações propostas individualmente referem-se à alteração de dois resíduos, porém, em alguns casos, um dos resíduos propostos já existe na proteína naquele local, implicando numa mutação de um único resíduo, afetando menos a cadeia de resíduos da proteína nativa, podendo ser uma melhor opção, de acordo com o interesse do usuário pesquisador.

Neste trabalho foram apresentados os principais aspectos da construção do banco de dados bem como da ferramenta web para acessá-lo e alguns resultados usando a interação ponte dissulfeto. A busca e análise dos resultados abordando as outras interações se dão da mesma forma, tanto para a busca quanto para a avaliação dos possíveis mutantes.

Com relação a estes aspectos, o detalhe da conformação já foi abordado anteriormente mas, é interessante ressaltar a forma de pesquisa do banco, aqui proposta. Inicialmente, para a validação da conformação, uma busca baseada na sobreposição um a um foi realizada e demonstrou fornecer resultados interessantes, porém computacionalmente caros (em estimativas de tempo de execução e necessidade de poder de processamento). Assim, a proposta do uso do *score* se tornou mais viável computacionalmente além de mais precisa que a avaliação da maior distância obtida na sobreposição, uma vez que o *score* é baseado nas contribuições das várias distâncias calculadas para cada átomo, gerando parâmetros mais ricos de informação. O uso do *score* também permitiu, além da redução do número de sobreposições realizadas na busca, uma redução no número de arquivos gerados por estas sobreposições, pois para cada sobreposição um arquivo no formato PDB é gerado para que seja importado em ferramentas como o jmol, para visualização. O *score* gerado e apresentado considera a contribuição de cada uma dessas 12 distâncias na sobreposição, tornando-se uma assinatura ou representação daquela interação no banco permitindo inclusive ordenar as sobreposições a partir desse valor, gerando o vetor a ser pesquisado por busca binária.

Nesta busca, procura-se encontrar a menor diferença entre o valor do *score* do par da proteína-alvo e o valor do *score* do par da interação. Quanto menor a diferença, mais próxima é a conformação do par da proteína com o par da interação do banco encontrada. A vantagem do uso desse valor é que a busca a ser realizada no banco consiste em apenas uma consulta SQL[49] simples no banco hospedado no MySQL, sem a necessidade da sobreposição do par da proteína com todos os pares da interação para encontrar uma proximidade entre eles.

Estes resultados demonstram que o propósito do banco de dados de interações se confirma, pois as mutações propostas, por exemplo substituição dos resíduos ASN12A e TYR15A ambos por CYS, sugerem que a proteína aumente sua estabilidade, considerando o resultado apresentado pelas ferramentas de análise de mutação simples. Porém, não é escopo, do

[49] Structured Query Language – Linguagem de consulta utilizada em sistemas de gerenciamento de bancos de dados

trabalho aqui apresentado, realizar tal análise, ficando a cargo do usuário pesquisador. Da mesma forma, não se garante que toda mutação proposta alcance o aumento da estabilidade, ficando a cargo do usuário pesquisador avaliar qual mutação melhor lhe convém de acordo com suas necessidades e com a proteína envolvida.

6.1 Projetos futuros

A partir desse trabalho, novos projetos ou melhorias podem advir:

6.1.1 Diferença das distâncias

Usar a diferença das distâncias por átomo para refinar a busca do par da interação a partir do par da proteína alvo. Considerando cada valor de delta obtido (12) no arquivo gerado para a proteína alvo, refinar a busca do par da interação avaliando cada um desses valores do arquivo da proteína, comparados com os respectivos valores do arquivo da interação. O par da interação que apresentar as menores diferenças para os 12 valores será o mais próximo ou um dos mais aproximados para a sobreposição.

6.1.2 10 referências

Cada interação é caracterizada por suas 16 distâncias e pelo arquivo de referência (maior resolução alta) usado na sobreposição. Ao invés de ter somente esta referência, usar outras 9, distribuídas linearmente a partir dessa primeira, de forma que sejam feitas 10 sobreposições ao invés de uma só para cada par da proteína alvo. Esta estratégia possibilitará ter várias referências e a que será usada será a mais próxima do par usado na busca, favorecendo a pesquisa de valores baixos de score, ou seja, com menores contribuições das diferenças das posições atômicas resultando em buscas mais eficientes do ponto de vista da precisão, porém com um gasto computacional um pouco maior.

6.1.3 Avaliação de estabilidade e atividade

Uma integração interessante que pode ser realizada é com ferramenta de avaliação de estabilidade ou desenvolvimento de módulo para isto, avaliando o par a ser mutado mais a estabilidade proporcionada pela mutação dos dois resíduos concomitantemente. Isto proporciona mais informações sobre o par sugerido para mutação, apesar de que a avaliação da estabilidade realizada por estas ferramentas refere-se a mutações de um único resíduo de aminoácido.

Integração com ferramenta de avaliação de atividade da proteína ou desenvolvimento de módulo para isto, permitindo que a análise do par a ser mutado, além da avaliação da manutenção da estabilidade permita a avaliação da manutenção da atividade da proteína em questão.

6.1.4 Validação em bancada

Para confirmar a eficácia do método, há a necessidade da validação das mutações propostas através da expressão da proteína mutante em laboratório. Além disso, a realização de testes de atividade nas proteínas mutantes podem complementar a validação confirmando a manutenção da atividade proteica após a mutação.

6.1.5 Alterações na interface

Substituir a visualização de três pares (*previous*, *target*, *matched*, como observados na Figura 44) pela sobreposição dos pares da interação (em uma cor) com os pares da proteína alvo (em outra cor), porém, com a cadeia lateral desses, de forma a se perceber visualmente as características da sobreposição. A exibição dessa sobreposição implica na geração dos novos arquivos a serem sobrepostos (contendo a cadeia lateral) além do próprio processo de sobreposição, gerando novo arquivo. Este processo deve ocorrer apenas após a definição de um grupo de candidatos para que os arquivos sejam gerados apenas para estes. Assim, o custo computacional pode ser reduzido.

6.1.6 Support Vector Machine

Outro projeto interessante é a verificação da possibilidade do uso de *Support Vector Machine* para a predição com um conjunto de treinamento baseado no banco de dados criado possibilitando a predição das interações com base neste conjunto. O uso do SVM poderá agilizar o processo de busca do par da interação ou do grupo candidato, havendo necessidade de validação visual (sobreposição) para a confirmação do par escolhido.

6.1.7 Interações entre cadeias

Ignoradas inicialmente para a montagem da base de dados, as interações entre cadeias podem e devem ser incorporadas no banco de interações. Sua remoção dos parâmetros iniciais permitiu a definição clara da busca e da geração dos arquivos. Sua adição neste momento envolve alteração de código e interface para permitir ao usuário filtrar este tipo de interação na busca a ser realizada no banco dando-lhe mais opções e liberdade de escolha. Dessa forma, ao escolher uma interação o usuário poderá definir se ela deve ocorrer entre cadeias, numa mesma cadeia ou em ambos os casos.

Referências

Almog O., Gallagher D. T., Ladner J. E., Strausberg S., Alexander P., Bryan P., Gilliland G. L. Structural basis of thermostability. Analysis of stabilizing mutations in Subtilisin BPN'. *J Biol Chem*. 2002 Jul 26;277(30):27553-8. Epub 2002 May 13.

Alva V., Syamala D.D.P., Sowdhamini R. COILCHECK: an interactive server for the analysis of interface regions in coiled coils. *Protein Pept Lett*. 2008;15(1):33-8. PMID: 18221010.

Baase, W.A., Liu, L., Tronrud, D.E., Matthews, B.W. Lessons from the lysozyme of phage T4. *Protein Science*, 19: 631–641. 2010, doi: 10.1002/pro.344.

Baker E.N., Hubbard R.E. 1984. Hydrogen Bonding in Globular Proteins. *Prog. Biophys. Molec. Biol.* 44, 97-179.

Berg, J. M.; Tymoczko, J.L; Stryer, L. *Biochemistry*, 6.ed. New Jersey: John Wiley, 2006.

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig, H. Shindyalov I.N., Bourne P.E. The Protein Data Bank. URL: http://www.pdb.org. *Nucleic Acids Research*, 28: 235-242, 2000.

Berndt K.D., Güntert P., Wüthrich K. The NMR Structure of the Dendrotoxin K from the Venom of Dedroaspis polylepis polylepis in Solution. *J. Mol. Biol.*, 234, 735-750, 1993.

Betts M.J., Russell R.B. Amino acid properties and consequences of subsitutions. In *Bioinformatics for Geneticists*, M.R. Barnes, I.C. Gray eds, Wiley, 2003. Disponível em: <http://www.russelllab.org/aas/>. Acessado em: 20/09/2012.

Betz, S.F. Disulfide bonds and the stability of globular proteins. *Protein Sci.*, 2, 1551-1558, 1993.

Bhattacharyya R., Pal D., Chakrabarti P. Disulfide bonds, their stereospecific environment and conservation in protein structures. *Protein, Engineering, Design & Selection*, vol. 17 no. 11 pp. 795–808, 2004. Published online December 2, 2004 doi:10.1093/protein/gzh093.

Boisbouvier J., Blackledge M., Sollier A., Marion D. Simultaneous determination of disulphide bridge topology and three-dimensional structure using ambiguous intersulphur distance restraints: Possibilities and limitations. *Journal of Biomolecular NMR*. Springer Netherlands, Issn: 0925-2738, pp 197-208, Vol 16, n 3, Doi: 10.1023/A:1008354007926, 2000.

Brevern, A.G. *Venn diagram of Amino Acids*. French Institute of Health and Medical Research. 10/11/2006. Disponível em: <http://www.dsimb.inserm.fr/~debrevern/VENN_DIAGRAM>. Acessado em: 18/12/2011.

Carugo, O. Statistical validation of the root-mean-square-distance, a measure of protein structural proximity. *Protein Engineering, Design & Selection*, vol. 20 no. 1 pp 33-37, (2007) 20 (1): 33-37. doi: 10.1093/protein/gzl051 First published online: January 11, 2007.

Ceroni A., Passerini A., Vullo A., Fraconi P. DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Research*, 2006, vol. 34, web server issue: w177-w181.

Cheng J., Randall A., Baldi P. Prediction of Protein Stability Changes for Single-Site Mutations Using Support Vector Machines. *Proteins*, vol 62, no. 4, pp. 1125-1132, 2006.

Cheng J., Saigo H., Baldi P. Large-Scale Prediction of Disulphide Bridges Using Kernel Methods, Two-Dimensional Recursive Neural Networks, and Weighted Graph Matching. *Proteins: Structure, Function, Bioinformatics*, vol 62, no. 3, pp. 617-629, 2006.

Cohen M, Potapov V, Schreiber G. Four distances between pairs of amino acids provide a precise description of their interaction. PLoS Computational Biology. 2009 Aug;5(8):e1000470. Epub 2009 Aug 14.

Deutsch C., Krishnamoorthy B. Four-Body Scoring Function for Mutagenesis. *Bioinformatics*, 2007 23(22):3009-3015; doi:10.1093/bioinformatics/btm481.

Dill, K.A. Dominant forces in protein folding. *Biochemistry*, Volume 29, issue 31 (1990), p. 7133-7155. ISSN: 0006-2960 DOI: 10.1021/bi00483a001. American Chemical Society.

Donaldson Jr. S.H., Lee Jr. C.T., Chmelka, B.F., Israelchvili J.N. General hydrophobic interaction potential for surfactant/lipid bilayers from direct force measurements between light-modulated bilayers. *Proceedings of the National Academy of Sciences of USA - PNAS* 2011 Vol 108, num 38, 15699-15704; published ahead of print September 6, 2011, doi: 10.1073/pnas.1112411108.

Eftink M., Pedigo S. Protein Folding, In: Editor-in-Chief: Robert A. Meyers, Editor(s)-in-Chief, *Encyclopedia of Physical Science and Technology* (Third Edition), Academic Press, New York, 2003, Pages 179-190, ISBN 9780122274107, 10.1016/B0-12-227410-5/00614-1.

Elmasri R., Navathe, S.B. Fundamentals of Database Systems, 3rd ed., Addison-Wesley, MA, 2000.

Eswar N., Ramakrishnan C. Deterministic features of side-chain main-chain hydrogen bonds in globular protein structures. *Protein Engineering Design & Selection - PEDS*. (2000) 13(4): 227-238 doi:10.1093/protein/13.4.227

Ferrè F., Clote P. DiANNA: A web server for disulfide connectivity prediction. *Nucleic Acids Research*, 2005b; 33: 230-2.

Ferrè F., Clote P. Disulfide connectivity prediction using secondary structure information and diresidue frequencies. *Bioinformatics*, May 15, 2005a; 21(10): 2336-2346.

Franks, F. Protein stability: the value of 'old literature', *Biophysical Chemistry*, Volume 96, Issues 2–3, 2 May 2002, Pages 117-127, ISSN 0301-4622, 10.1016/S0301-4622(02)00014-5.

González-Díaz H., Molina R., Uriarte E. Recognition of stable protein mutants with 3D stochastic average electrostatic potentials, *FEBS Letters*, Volume 579, Issue 20, 15 August 2005, Pages 4297-4301, ISSN 0014-5793, 10.1016/j.febslet.2005.06.065.

Gromiha, M.M. Chapter 6 - Protein Stability, *Protein Bioinformatics*, Academic Press, Singapore, 2010, Pages 209-245, ISBN 9788131222973, 10.1016/B978-8-1312-2297-3.50006-0.

Gromiha, M.M.; Selvaraj, S. Inter-residue interactions in protein folding and stability. *Progress in Biophysics and Molecular Biology*, Volume 86, Issue 2, October 2004, Pages 235-277, ISSN 0079-6107, 10.1016/j.pbiomolbio.2003.09.003.

Guerois R., Nielsen J.E., Serrano L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *Journal of Molecular Biology*, Volume 320, Issue 2, 5 July 2002, Pages 369-387, ISSN 0022-2836, 10.1016/S0022-2836(02)00442-4.

Harris R.C., Bredenberg J.H., Silalahi A.R.J., Boschitsch A.H., Fenley M.O. Understanding the physical basis of the salt dependence of the electrostatic binding free energy of mutated charged ligand–nucleic acid complexes, *Biophysical Chemistry*, Volume 156, Issue 1, June 2011, Pages 79-87, ISSN 0301-4622, 10.1016/j.bpc.2011.02.010.

Hazes B., Dijkstra B.W. Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein Engineering* 2: 119-125, (1988).

Hinz H.-J., Steif C., Vogl T., Meyer R., Renner M., and Ledermuller R. Fundamentals of protein stability. *Pure and Applied Chemistry*, 1993, Vol. 65, No. 5, pp. 947-952, 1993.

Huang L.T., Saraboji k., Ho S.Y., Hwang S.F., Ponnuswamy M.N., Gromiha M.M. Prediction of protein mutant stability using classification and regression tool, *Biophysical Chemistry*, Volume 125, Issues 2–3, February 2007, Pages 462-470, ISSN 0301-4622, 10.1016/j.bpc.2006.10.009.

Hunter, L. Molecular biology for computer scientists. In *Artificial intelligence and Molecular Biology*, L. Hunter, Ed. American Association for Artificial Intelligence, Menlo Park, CA, 1-46. 1993.

Jackson, S.E. Protein Folding, Engineering of, In: Editors-in-Chief: Franco Bassani, Gerald L. Liedl, and Peter Wyder, Editor(s)-in-Chief, *Encyclopedia of Condensed Matter Physics*, Elsevier, Oxford, 2005, Pages 418-425, ISBN 9780123694010, 10.1016/B0-12-369401-9/00382-X.

Jargas, A.M. Shell Script Profissional. Ed. Novatec, São Paulo, 2008.

Jeffrey G.A. *An Introduction to Hydrogen Bonding*. 1997. Oxford University Press.

Kabsch W., Sander C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* 22, 2577-2637. 1983.

Knuth D. "Section 6.2.1: Searching an Ordered Table". Sorting and Searching. *The Art of Computer Programming.* 3rd ed. Addison-Wesley. pp. 409–426. ISBN 0-201-89685-0. 1997.

Krasil'nikov, P. M., Pashchenko, V. Z., Noks, P. P., and Corresponding Member of the RAS A. B. *Rubin Doklady Biochemistry and Biophysics*, Vol. 376, 2001, pp. 16–18. Translated from Doklady Akademii Nauk, Vol. 376, No. 3, 2001, pp. 404–406.

Kuroki, R., Weaver, L.H., Matthews, B.W. A covalent enzyme-substrate intermediate with saccharide distortion in a mutant T4 lysozyme. *Science*, 24 December 1993: 262 (5142), 2030-2033. [DOI:10.1126/science.8266098]

Kuroki, R., Weaver, L.H., Matthews, B.W. Structural basis of the conversion of T4 lysozyme into a transglycosidase by reengineering the active site. *Proc. Natl. Acad. Sci. USA Biochemistry*, Vol. 96, pp. 8949–8954, August 1999.

Laskowski, R.A., Moss, D.S., Thornton, J.M. Main-Chain bond lengths and bond angles in protein structures. *Journal of Molecular Biology*. (1993) 231, 1049-1067.

Lee F.S., Warshel A. A local reaction field method for fast evaluation of longrange electrostatic interaction in molecular simulations. *Journal of Chemical Physics*. 1997, 3100 (1992); doi: 10.1063/1.462997.

Lehninger A. L., Nelson D. L., Cox M. M. *Princípios de Bioquímica.* 4ª Edição. Editora: Sarvier, 1232p, 2007.

Lesser G.J., Rose G.D. Hydrophobicity of amino acid subgroups in proteins. *Proteins: Struct. Funct. Genet*. 8, 6-13,1990.

Lin H., Tseng L. DBCP: a web server for disulfide bonding connectivity pattern prediction without the prior knowledge of the bonding state of cysteines. *Nucleic Acids Research*, 2010, vol. 38, no. suppl_2 w503-w507.

Lins L., Brasseur R. The hydrophobic effect in protein folding. *The Journal of the Federation of American Societies for Experimental Biology – FASEB J*. April 1995, Vol 9, n. 7, 535-540.

Livingstone, C.D.; Barton, G.J. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci*, 1993. 9(6): 745-756 doi:10.1093/bioinformatics/9.6.745.

Mancini A.L., Higa R.H., Falcão P.R.K., Yamagishi M.E.B., Neshich G. Cálculo de possíveis contatos atômicos internos a uma proteína ou entre proteínas no software SMS. Campinas: Embrapa Informática Agropecuária, dez, 2003. 8 p. *Embrapa Informática Agropecuária. Comunicado Técnico 59*. Disponível em: <http://www.infoteca.cnptia.embrapa.br/bitstream/doc/8835/1/comtec59.pdf>. Acessado em: 10/04/2012. ISSN 1677-8464.

Magliery T.J., Lavinder J.J., Sullivan B.J. Protein stability by number: high-throughput and statistical approaches to one of protein science's most difficult problems. *Current Opinion in Chemical Biology*, Volume 15, Issue 3, June 2011, Pages 443-451

Martin, A.J.M.; Vidotto, M.; Boscariol, F.; Di Domenico, T.; Walsh, I.; Tosatto, S.C.E. RING: Networking interacting residues, evolutionary information and energetics in protein structures. *Bioinformatics*, 2011 Apr 14.

Mason, J.M., Bendall, D.S., Howe, C.J., Worrall, J.A.R. The role of a disulfide bridge in the stability and folding kinetics of *Arabidopsis thaliana* cytochrome c_{6A}. *Biochimica et Biophysica Acta* 1824, 2012, 311-318 doi:10.1016/j.bbapap.2011.10.015.

Masso, M. & Vaisman, I.I. AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements, 2010, *Protein Eng. Des. Sel.* **23**, 683-687.

Matsumara, M., Becktel, W.J., Levitt, M., Matthews, B.W. Stabilization of phage T4 lysozyme by engineered disulfide bonds. *Proc. Natl. Acad. Sci. USA*, 86, 6562-6566, 1989.

McDonald I., Thornton J.M. Atlas of Side-Chain and Main-Chain Hydrogen Bonding. Web edition 1994. Original edition 1993. Disponível em: <http://www.biochem.ucl.ac.uk/bsm/atlas>. Acessado em: 07/08/2012.

Miyawaki O., Tatsuno M. Thermodynamic analysis of alcohol effect on thermal stability of proteins. *Journal of Bioscience and Bioengineering*, Volume 111, Issue 2, February 2011, Pages 198-203, ISSN 1389-1723, 10.1016/j.jbiosc.2010.09.007.

Mizuguchi K., Deane C.M., Blundell T.L., Johnson M.S., Overington J.P. JOY: protein sequence-structure representation and analysis. *Bioinformatics*. 1998;14:617-623.

Morris A.L., MacArthur M.W., Hutchinson, E.G., Thornton, J.M. Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Bioinformatics*. Vol 12, Num 4, p 345-364. Wiley Subscription Services, Inc., A Wiley Company. SN 1097-0134, DOI 10.1002/prot.340120407.

Motta, V.T. *Bioquímica*. Ed. Medbook. 2ª edição, 2011. 488pp. ISBN 8599977660.

Muskal S., Holbrook S., Kim S. Prediction of the disulfide-bonding state of cysteine in proteins. *Protein Engineering*, vol. 3, no. 8, pp. 667-672, 1990.

Nakamoto K., Margoshes M., Rundle R.E. Stretching Frequencies as a Function of Distances in Hydrogen Bonds. *Journal of the American Chemical Society*, 1955**,** *77* (24), 6480-6486. DOI**:** 10.1021/ja01629a013.

Neves, J.C. Programação Shell Linux. Brasport, 6ª Ed, Rio de Janeiro, 2006, ISBN 85-7452-264-3.

Osherovich, L. Engineering protein stability. *Nature - SciBX: Science–Business eXchange.* Volume 4, Number 7. Nature Publishing Group. Published online Feb. 17 2011; doi:10.1038/scibx.2011.184. Disponível em: <http://www.nature.com/scibx/journal/v4/n7/pdf/scibx.2011.184.pdf>. Acessado em: 07/07/2012.

Pace C.N., Shirley B.A., McNutt M., Gajiwala K. Forces contributing to the conformational stability of proteins. *The journal of the Federation of American Societies for Experimental Biology - FASEB J*. January 1996 10:75-83

Pace, C. N. Evaluating the contribution of hydrogen bonding and hydrophobic bonding to protein folding. *Methods Enzymol*. 259, 538-554, 1995.

Pace, C.N. Measuring and increasing protein stability, *Trends in Biotechnology*, Volume 8, 1990, Pages 93-98, ISSN 0167-7799, 10.1016/0167-7799(90)90146-O.

Pace, C.N., Fu, H., Fryar, K.L., Landua, J., Trevino, S.R., Shirley, B.A., Hendricks, M.M., Iimura, S., Gajiwala, K., Scholtz, J.M., Grimsley, G.R. Contribution of Hydrophobic Interactions to Protein Stability, *Journal of Molecular Biology*, Volume 408, Issue 3, 6 May 2011, Pages 514-528, ISSN 0022-2836, 10.1016/j.jmb.2011.02.053.

Pace, C.N., Horn, G., Hebert, E.J., Bechert, J., Shaw, K., Urbanikova, L., Scholtz, J.M., Sevcik, J. Tyrosine hydrogen bonds make a large contribution to protein stability, *Journal of Molecular Biology*, Volume 312, Issue 2, 14 September 2001, Pages 393-404, ISSN 0022-2836, 10.1006/jmbi.2001.4956.

Pantoliano M. W., Ladner R. C., Bryan P. N., Rollence M. L., Wood J. F., Poulos T. L. Protein engineering of Subtilisin BPN': enhanced stabilization through the introduction of two cysteines to form a disulfide bond. *Biochemistry*. 1987 Apr 21;26(8):2077-82. PMID: 3476160 [PubMed - indexed for MEDLINE].

Pantoliano M. W., Whitlow M., Wood J. F., Dodd S. W., Hardman K. D., Rollence M. L., Bryan P. N. Large increases in general stability for Subtilisin BPN' through incremental changes in the free energy of unfolding. *Biochemistry*. 1989 Sep 5;28(18):7205-13. PMID: 2684274 [PubMed - indexed for MEDLINE].

Pantoliano M. W., Whitlow M., Wood J. F., Rollence M. L., Finzel B. C., Gilliland G. L., Poulos T. L., Bryan P. N. The engineering of binding affinity at metal ion binding sites for the stabilization of proteins: Subtilisin as a test case. *Biochemistry*. 1988 Nov 1;27(22):8311-7. PMID: 3072018 [PubMed - indexed for MEDLINE].

Pereira H. M., Franco G. R., Cleasby A., Garratt R. C. Structures for the Potential Drug Target Purine Nucleoside Phsophorylase From Schistosoma mansoni Casual Agent of Schistosomiasis. *Journal of Molecular Biology*, Amsterdam, v. 353, p. 584-599, 2005.

Perutz, M.F. Electrostatic effects in proteins. *Science*, Sep 29, 201 (4362), 1187–1191, 1978.

Petersen, M.T., Jonson, P.H., Petersen, S. B. Amino acid neighbours and detailed conformational analysis of cysteines in proteins. *Protein Engineering*. 12, 535–548, 1999.

Richardson J.S. The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry.* 1981; Vol 34: 167-339. PMID: 7020376. ISBN 0-12-034234-0.

Rose G.D., Fleming P.J., Banavar J.R., Maritan A. A backbone-based theory of protein folding. *Proceedings of the National Academy of Sciences of United States of America - PNAS*, 2006 103 (45) 16623-16633; published ahead of print October 30, 2006, doi:10.1073/pnas.0606843103.

Sakaguchi M., Matsuzaki M., Niimiya K., Seino J., Sugahara Y., Kawakita M. Role of proline residues in conferring thermostability on aqualysin I. *J Biochem.* 2007 Feb;141(2):213-20. Epub 2006 Dec 14.

Sakaguchi M., Takezawa M., Nakazawa R., Nozawa K., Kusakawa T., Nagasawa T., Sugahara Y., Kawakita M. Role of Disulphide Bonds in a Thermophilic Serine Protease Aqualysin I from Thermus aquaticus YT-1. *J Biochem.* 2008 May;143(5):625-32. Epub 2008 Jan 23.

Singh J., Thornton J.M. *Atlas of Protein Side-Chain Interactions, Vols. I & II*, IRL press, Oxford, 1992. Dataset disponível em: <http://www.biochem.ucl.ac.uk/bsm/sidechains>. Acessado em: 15/01/2011.

Singh R. A review of algorithmic techniques for disulfide-bond determination. *Briefings in Functional Genomics and Proteomics*, March 27, 2008, 1-16.

Sowdhamini R., Srinivasan N., Shoichet B., Santi D., Ramakrishnan C., Balaram P. Stereochemical modeling of disulfide bridges. Criteria for introduction into proteins by site-directed mutagenesis. *Protein Engineering*, vol. 3, no. 2, pp. 95-103, 1989.

SSBOND. 1999. Disponível em: <http://eagle.mmid.med.ualberta.ca/forms/ssbond.html>. Acessado em: 20/07/08.

Takahashi T. Significant role of electrostatic interactions for stabilization of protein assemblies, *Advances in Biophysics*, Volume 34, 1997, Pages 41-54, ISSN 0065-227X, 10.1016/S0065-227X(97)89630-X.

Taylor R., Kennard O., Versichel W. The geometry of the N-H···O=C hydrogen bond. 3. Hydrogen-bond distances and angles. *Acta Crystallographica Section B*. Vol 40, Part 3, 280-288, 1984. doi: 10.1107/S010876818400210X.

Taylor W. R. The Classification of Amino Acid Conservation. *J. Theor. Biol.* 119, 205-218. 1986.

Teilum K., Olsen J.G., Kragelund B.B. Protein stability, flexibility and function. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, Volume 1814, Issue 8, August 2011, Pages 969-976.

Tigerström A., Schwarz F., Karlsson G., Okvist M., Alvarez-Rua C., Maeder D., Robb F. T., Sjolin L. Effects of a novel disulfide bond and engineered electrostatic interactions on the thermostability of azurin. *Biochemistry* 43:12563-12574, (2004).

Tormo J., Stadler E., Skern T., Auer H., Kanzler O., Betzel C., Blaas D., Fita I. Three-dimensional structure of the Fab fragment of a neutralizing antibody to human rhinovirus serotype 2. *Protein Sci.* 1**:** 1154-1161, 1992. PDB ID: 1bbd. DOI:10.2210/pdb1bbd/pdb.

Tsai C., Chen B., Chan C., Liu H., Kao C. Improving disulfide connectivity prediction with sequential distance between oxidized cysteines. *Bioinformatics Advanced Access* published on October 13, 2005, DOI 10.1093/bioinformatics/bti715.

Voet D., Voet J.G. *Biochemistry*. Fourth edition. John Wiley & Sons, inc. 2011 - 1248 pp.

Vogl T., Brengelmann R., Hinz H. J., Scharf M., Lötzbeyer M., Engels J. W. Mechanism of protein stabilization by disulfide bridges: calorimetric unfolding studies on disulfide-deficient mutants of the alpha-amylase inhibitor tendamistat. *Journal of Molecular Biology.* 1995 Dec 1;254(3):481-96.

Wallwork S.C. Hydrogen-Bond Radii. *Acta Crystallographica.* Vol 15, Part 8, august 62, 758-759, 1962. doi:10.1107/S0365110X6200198X.

Wells, J.A., Powers, D.B. In vivo formation and stability of engineered disulfide bonds in Subtilisin. *J Biol Chem.* 15, 6564–6570, 1986.

Werfhorst M.V.D. *UCSB Researchers First to Develop Equation that Predicts Molecular Forces in Hydrophobic Interaction.* University of California Santa Barbara. College of Engineering. News Release, october 11, 2011. Disponível em: <http://engineering.ucsb.edu/news/520>. Acessado em: 10/10/2012.