Ministério da Educação
Universidade Federal de Itajubá
Criada pela Lei nº 10.435, de 24 de abril de 2002

Universidade Federal de Itajubá

Caracterização da descontinuidade de fitas em favor de hélices em estruturas proteicas toda-beta

PIBIC

RELATÓRIO FINAL
UNIFEI
2012/2013

Aluno: **Matheus Venturyne Xavier Ferreira**
Matrícula: 21197
Curso: Engenharia da Computação, campus Itabira-MG

Fase/Período: 5º

Vigência: Março de 2012 a Fevereiro de 2013

Resumo

Tanto a confecção de novos fármacos quanto a identificação de novos alvos moleculares terapêuticos dependem de uma maior compreensão do mecanismo de ação (função) das biomoléculas, em especial das proteínas. Essa funcionalidade, por sua vez, está intimamente ligada a um maior entendimento dos aspectos estruturais da cadeia proteica. Existe, pois, uma real demanda por uma maior e melhor determinação funcional e estrutural das proteínas. Anfinsen, no início da década de 1970, demonstrou que toda a informação que uma proteína globular de baixo peso molecular precisava para assumir sua forma enovelada tridimensional estava contida na sequência primária de seus resíduos. Como faz uma proteína globular para enovelar-se em sua estrutura terciária valendo-se apenas da informação contida em sua sequência de aminoácidos é o que ficou conhecido como Problema do Enovelamento de Proteínas (PEP). O PEP constitui um dos maiores problemas em aberto da ciência moderna. Certamente, que uma das formas de atacar o PEP passa por uma maior compreensão dos mecanismos que levam as cadeias proteicas a se organizarem nas estruturas secundárias helicoidais e em fitas previstas por Pauling e Corey. Durante este projeto de iniciação científica, nós vislumbramos uma oportunidade (o objetivo geral) de identificar e caracterizar possíveis padrões de propensão nas transições entre hélices e fitas em estruturas que são preponderantemente compostas por fitas.

Palavras-chave: bioinformática; proteínas; relação sequência-estrutura.

Sumário

I - Introdução....4
II - Objetivos....7
III - Descrição das atividades desenvolvidas....7
IV - Resultados obtidos e análise dos mesmos....12
V - Conclusões e metas futuras....22
VI - Metodologia e Plano de Trabalho....23
VII - Cronograma de Atividades....24
VI - Referências Bibliográficas....24
APÊNDICE A....28
APÊNDICE B....30
APÊNDICE C....31

I - Introdução

A bioinformática emergiu com uma nova e interdisciplinar ciência com o objetivo explícito de tentar organizar, minerar e correlacionar padrões complexos oriundos da gigantesca profusão de dados biológicos que vem inundando os periódicos e repositórios de dados desde a década de 1990. Divide-se, hoje, em pelo menos três grandes sub-áreas: bioinformática genômica, bioinformática estrutural e bioinformática funcional. A bioinformática estrutural tem como foco o estudo preditivo e comparativo de dados funcionais e estruturais de biomoléculas.

As proteínas[1] são uma das mais importantes moléculas dos seres vivos. Elas estão envolvidas em uma ampla gama de processos bioquímicos: nos componentes estruturais, nas reações enzimáticas; na contração muscular, movimento ciliar, flagelar, deformação e divisão celular; nas respostas imunológicas; na auto-reconstituição e reparação de tecidos; na regulação hormonal; no transporte de substâncias vitais no sangue e transporte celular inter membrana; no impulso nervoso; na reserva e armazenagem de nutrientes. Não é surpresa então que sejam a segunda substância mais encontrada nos seres vivos: 15% da massa de uma célula é proteína [2].

Toda essa espantosa diversidade de funções bioquímicas é feita pela combinação de 20 tipos de unidades monoméricas, chamados aminoácidos [2]. No processo de polimerização, os aminoácidos se ligam linearmente uns aos outros podendo formar extensas cadeias peptídicas (a estrutura primária). O número de proteínas sequencialmente diferentes para uma cadeia de 100 resíduos parece ser mais que suficiente para dar conta de toda essa diversidade funcional: 20100 ou 10130 combinações. As ligações flexíveis carbono-carbono e carbono-nitrogênio ao longo da cadeia principal conferem a uma boa parte dessas sequências a capacidade de existirem ainda sob uma grande quantidade de arranjos estruturais tridimensionais ou conformações: algo da ordem de 1050 possibilidades[2], para uma molécula de 100 resíduos [3].

Os primeiros estudos cristalográficos apontavam para uma relação estreita entre função e estrutura em proteínas. Apenas um conjunto mínimo de conformações dentre as 10^{100} deveria reunir as condições topológicas e químicas necessárias para torná-las funcionais. A própria capacidade das proteínas formarem cristais indicava a existência de uma conformação única e mais rígida [2]. Na década 1930, Linus Pauling e Robert Corey previram[3], com base em dados cristalográficos, a existência de refinados padrões conformacionais ou estruturas secundárias, como as hélices, voltas e fitas [4]. Em proteínas globulares estas estruturas poderiam dobrar sobre si mesmas (enovelar), escondendo

[1] A palavra “proteína”, como também “próton”, vem do étimo “protêios”, radical grego que significa primeiro ou o mais importante, também usado como prefixo indicando antecedência, como em “protótipo” e “protozoário” [1].

[2] Assumindo 3 graus de liberdade para cada ligação peptídica, na composição de hélices, voltas ou fitas, ou $3100 \simeq 1050$.

[3] Previsão confirmada 30 anos depois, pelas primeiras proteínas com estrutura 3D resolvidas [5].

resíduos hidrofóbicos no interior da cadeia e formando intricados arranjos tridimensionais ou uma estrutura terciária (figura 01).

Mas experimentos termodinâmicos já mostravam que as proteínas não eram tão organizadas assim. A estabilidade termodinâmica de proteínas globulares e de baixo peso molecular é marginal, da ordem de 5 a 20 Kcal mol^{-1} [7]. Essas proteínas desenovelam-se e perdem sua função sob ação de agentes (tais como calor, pressão, pH e osmólitos) que tenham capacidade de intervir nas fracas interações que mantêm sua estrutura nativa [8]. Mutações, envolvendo mudanças na composição ou na ordem de seus resíduos também interferem quase sempre negativamente em sua estabilidade termodinâmica [9]. Nem rígidas nem disformes, proteínas globulares são moléculas que vivem na borda do caos.

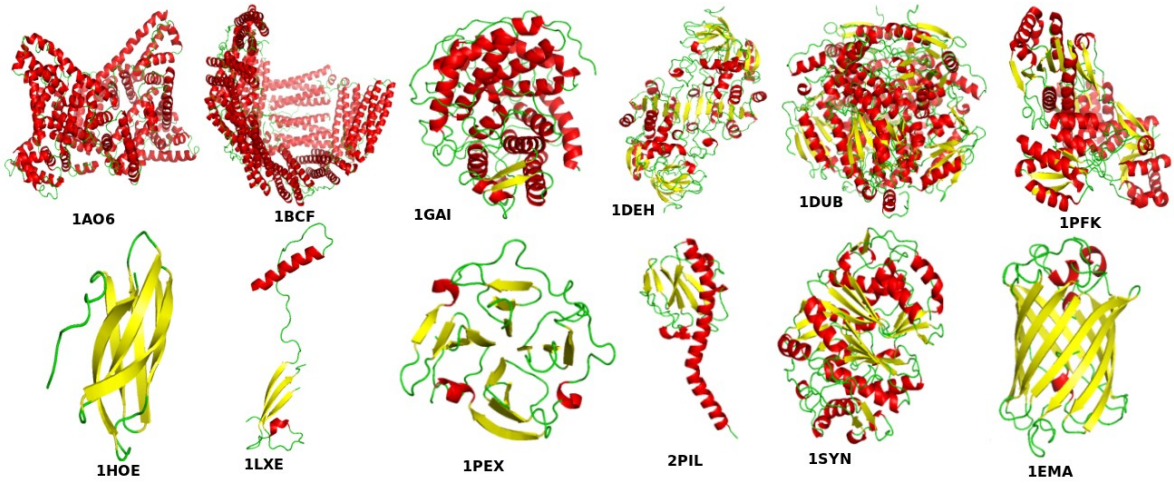


Figura 01: Exemplos do intrincado arranjo entre hélices e fitas comumente verificado em proteínas. Os códigos abaixo indicam os identificadores do PDB - Protein Data Bank [19].

Se é realmente uma determinada conformação que dá especificidade funcional a uma proteína, como é que essa estrutura é montada ou enovelada? Levinthal em 1967 já ponderava que esse processo de enovelamento não poderia envolver todas as virtualmente infinitas possibilidades estruturais que uma proteína poderia assumir em seu imenso espaço conformacional, e que certamente deveria haver heurísticas para indicar quais caminhos ou atalhos levariam ao enovelamento em sua forma funcional [10].

Foi então na década de 1970 que Christian Anfinsen e seus colegas descobriram que se fossem removidos os agentes desnaturantes e redutores que haviam levado ribonucleases *in vitro* a desenovelarem-se em conformações aleatórias e não funcionais, elas rapidamente retomavam sua configuração funcional nativa. Ficava claro que toda a informação que uma proteína globular de baixo peso molecular precisava para assumir sua forma enovelada tridimensional estava contida na sequência primária de seus resíduos[1] [11].

[1] In vivo, o enovelamento, principalmente de proteínas maiores e com peso molecular acima de 20 kDa, parece contar com o auxílio de outras proteínas, chamadas Chaperones,[20].

Como faz uma proteína globular para enovelar-se em sua estrutura terciária valendo-se apenas da informação contida em sua estrutura primária (sequência de aminoácidos) é o que ficou conhecido como *Problema do Enovelamento de Proteínas* (PEP) [12]. O PEP constitui um dos maiores problemas em aberto da ciência moderna. Sua complexidade está relacionada ao fato de haver uma degeneração informacional causada pela alta redundância do espaço sequencial em relação ao espaço conformacional: como vimos, para proteínas de 100 resíduos, 10^{130} sequências colapsam em 10^{50} conformações nativas; ou seja, em teoria, 1080 sequências podem produzir um mesmo padrão estrutural [12].

Certamente, que uma das formas de atacar o PEP passa por uma maior compreensão dos mecanismos que levam as cadeias proteicas a se organizarem nas estruturas secundárias helicoidais e em fitas previstas por Pauling e Corey [4]. Para este projeto de iniciação científica, nós vislumbramos uma oportunidade de identificar e caracterizar possíveis padrões de propensão nas transições entre hélices e fitas em estruturas que são preponderantemente compostas por fitas.

Um exemplo pode ser visto na figura 02. Temos ali um inibidor de metaloprotease composto essencialmente por um arranjo de fitas na forma de um barril. Mas, de forma um tanto quanto misteriosa, vemos que uma fita, de um conjunto de 3, faz uma transição meio abrupta por uma pequena hélice. Por que essa região em específico teve preferência por se organizar de forma diferente, helicoidal, interrompendo o padrão geral de fita? Que fatores entre composição de aminoácidos, interações, formas de empacotamento, tanto no contexto local como não-local, podem estar conspirando contra o padrão dominante de fitas em favor de uma organização helicoidal?

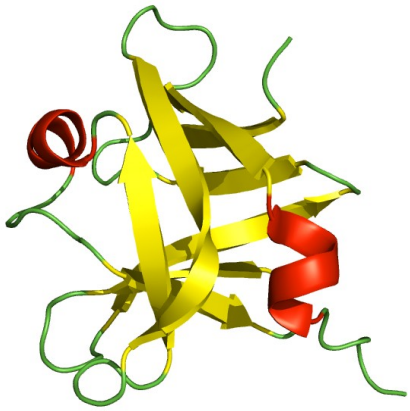


Figura 02: modelo estrutural da cadeia do inibidor de metaloproteinase de *Pseudomonas aeruginosa*, 1JIW:I. Percentagem de resíduos alfa de 12%, resíduos beta 51% e outros resíduos (laços) de 36% de um total de 105 resíduos. Dados baseados na atribuição de estrutura secundária pelo DSSP[23].

Essas são as perguntas que norteiam este projeto. Evidentemente que para respondê-las com alguma esperança de generalização, precisaremos juntar muito mais exemplos que o ilustrado na figura 02. O PDB - Protein Data Bank - é um repositório de estruturas

resolvidas, com coordenadas atômicas de várias biomoléculas, em sua maior parte proteínas [19]. Conta hoje com quase 77 mil estruturas catalogadas [22]. Faz parte deste projeto, um levantamento minucioso do PDB, filtrando estruturas que apresentem essa transição inusitada de fita em hélice, num banco de dados local. De posse das cadeias de interesse, usaremos técnicas computacionais diversas de mineração de dados, no intuito de extrair possíveis padrões que possam ajudar a identificar e caracterizar essas transições fitas-hélices.

Esse projeto conta com a parceria do programa de doutorado em Bioinformática da UFMG, do qual o proponente faz parte como co-orientador de diversas teses e projetos de pesquisa.

II - Objetivos

GERAIS: identificar e caracterizar regiões de descontinuidade de fitas em hélices em estruturas proteicas organizadas predominantemente por fitas.

ESPECÍFICOS:

- Contribuir para uma maior compreensão das relações entre sequência e estrutura no PEP.
- Montar um conjunto de scripts em PERL e R que auxiliem na filtragem das cadeias de interesse visando a montagem de um banco de dados local.
- Aprender e usar (ou mesmo criar) técnicas computacionais de mineração de dados adequadas a extração de padrões que possam auxiliar a identificar e caracterizar as transições fitas-hélices em cadeias com predominância de fitas.
- Propor novos algoritmos mais eficientes de identificação de estruturas secundárias, tendo em vista os insights que possam ser obtidos a partir da análise dos padrões nas descontinuidades de fitas em hélices.

III - Descrição das atividades desenvolvidas

- Estudo do formato de arquivo do PDB - Protein Data Bank
  PDB (Protein Data Bank) [19] é um repositório para os dados em 3-D de moléculas biológicas, tais como proteínas e ácidos nucleicos. Os dados geralmente são obtidos por cristalografia de raios X ou espectroscopia de RMN e registrados por pesquisadores do mundo todo, são livremente acessíveis na internet através dos sites de suas organizações membros. Dentre eles destacamos o PDB.org que pode ser acessado através do endereço www.pdb.org.

- Revisão de conceitos de bioquímica

Estudo de estrutura de proteínas, formas de agrupar seus diferentes grupos de acordo com sua estrutura [44]. Revisão bibliográfica sobre PEP [28], possíveis soluções e hipóteses para explicar o enovelamento de proteínas [30][40], se há um mecanismo unificado para o enovelamento [27] e as expectativas para a solução desse grande problema da atualidade [25]. Pesquisa sobre os métodos presentes de atribuição de estrutura secundária, sendo adotados até o momento DSSP [44], STRIDE [45], kaksi [41] e anotadores presentes no PDB.

- Desenvolvimento de scripts em PERL:
  Foi aproveitado um pacote de scripts perl para tratamento de arquivos PDB (PDBEST) [42]. Foram realizadas diversas alterações dentro do PDBEST, como adaptação dos scripts utilizados até agora para tratamento das cadeias proteicas utilizando paradigma de orientação a objetos, além de incorporação de novas funcionalidades. Até o momento há 20 scripts para a execução das tarefas de filtragem e processamento, 24 classes perl além de arquivos de configuração[1].

Inicialmente foi construída uma base de dados, fazendo um levantamento de quais IDs PDBs exemplos de descontinuidade ocorrem. Foi usado o *Advanced Search* do PDB, escolhendo SCOP [20] All Beta, com menos de 30% de identidade sequencial (retornou 2092 estruturas e foi adiciona mais uma proteína de interesse de código 1JIW).

Foi aplicado os scripts PERL para (PDBest):

- Classificar e identificar cada cadeia.
- Filtrar apenas as cadeias *All Beta*.
- Aplicar diferentes métodos de atribuição de estrutura secundária a cada cadeia.
- Extrair distâncias em resíduos entre estruturas secundárias (distâncias entre fitas, entre hélices e entre fitas com hélices).

Na tabela abaixo um resumo do processo de filtragem em cada script:

Script	Entrada	Saída
Pré-filtragem com *Advanced Search* do PDB, escolhendo `SCOP All Beta`, com menos de 30% de identidade sequencial.		1. Retornou 2092 estruturas e foi adicionada a proteína 1JIW que motivou esse trabalho devido sua descontinuidade de fita para hélice.
`filtra_pdb.pl`	2093 Pdb ids	1. 224 removidos por terem modelos.

[1] O download dos scripts até então gerados, junto às classes que utilizam, além da compilação de DSSP e STRIDE podem ser encontrados em https://www.dropbox.com/sh/ubr57a8acfnh28y/hZMWMMNulJ. Pode ser preciso instalar Kaksi no computador para utiliza-lo, as instruções para instalação em Linux é fornecida junto ao pacote kaksi.

		2. 1869 mantidos.
`separa_chain_pdb.pl`	1869 PDB IDs	1. 7696 cadeias
`removeDNA.pl`	7696 cadeias	1. 904 removidos por não terem aminoácidos. 2. 6792 mantidos.
`filtra_estrutura_secundaria.pl`	6792 cadeias	1. 236 cadeias removidas por não passarem no filtro de DSSP, Kaksi ou STRIDE (uma das condições é a cadeia conter ao menos 5 resíduos) 2. 6556 mantidos.
`uniao_estrutura.pl`	6556 cadeias	1. 6556 arquivos de descrição da estrutura secundária unindo os métodos de atribuição de estrutura secundária, como DSSP e STRIDE (processo de união é descrito no dia 10 de dezembro de 2011, diário 1).
`remove_chain.pl`	6556 cadeias	1. 683 cadeias removidas por terem 0% de resíduos de fitas (considerando a estrutura secundária gerada no passo anterior). 2. 5.873 mantidos.
`uniao_distancia.pl`	5873 descrições de SSE	7.234 arquivos de distâncias considerando cadeias com conflitos como cadeias diferentes.
`extrair_estruturas.pl`	5873 cadeias	Gerou 4 bases de dados, uma para transições (ou descontinuidades) de fitas para hélices, hélices para fitas, hélices para hélices e fitas para fitas.
`CSM.pl`	4 bases de dados de transições	Gerou matriz do CSM [52] para cada uma das bases de dados de transições.

Tabela 1: Quantidade de dados gerados por cada script e resumo do processo de filtragem na base de dados.

Cada script possui a seguinte função:

- `filtra_pdb.pl`

  Realiza uma filtragem na base de dados buscando por estruturas que possuem modelos. Estes estão relacionados ao método de resolução estrutural e envolveu uma técnica que gera “modelos”. Sendo comum quando se usa a RNM - Ressonância Nuclear Magnética - como ferramenta de resolução. Quando a técnica é a Cristalografia (de raios X ou mesmo de

Neutrons), geralmente não se trabalha com modelos. Estas estruturas foram guardadas e até o momento, concentramos nossa atenção nas resoluções por técnicas cristalográficas.

- `separa_chain_pdb.pl`
  Este segundo script foi empregado para separar as cadeias contidas na estrutura. Uma proteína pode ser formada por vários domínios/cadeias, compactos e segmentos fisicamente separados da cadeia polipeptídica [30].

- `removeDNA.pl`
  Foi empregado para remover cadeias de DNA, que não possuem aminoácidos na sequência.

- `filtra_estrutura_secundaria.pl`
  Neste script são utilizados os 3 métodos destacados de atribuição de estrutura secundária; DSSP, Kaksi, STRIDE. São compostos por programas que realizam sua própria filtragem das cadeias, caso haja alguma irregularidade essa cadeia é descartada da base de dados (normalmente cadeias que passam pelo removeDNA.pl por conter ao menos um resíduo mas que não chega a conter mais que 5 resíduos).

- `remove_chain.pl`
  Ferramenta para remover cadeias específicas da base de dados. Através de análises estatísticas utilizando códigos em linguagem R [46] da estrutura secundária fornecida pelo `filtra_estrutura_secundaria.pl` ou `uniao_estrutura.pl` pode-se identificar cadeias com 0% de resíduos beta segundo DSSP, STRIDE, kaksi, anotadores do PDB ou a sequência gerada pela união dos métodos. Essas cadeias foram removidas.

- `uniao_estrutura.pl`
  Esse script recupera a estrutura secundária atribuída para cada cadeia pelos diversos métodos citados junto com a anotação de estrutura do arquivo PDB e realiza uma união dessa atribuição. No entanto, podem ocorrer divergências de atribuição entre os métodos. Com isso define-se que a prioridade é que caso um método atribua uma fita ou hélice para um resíduo este deve ser usado. Caso diferentes métodos forneçam que um resíduo é fita e outro que corresponde a uma hélice este seria marcado como um conflito. Com isso sacrificamos precisão para conseguir uma maior cobertura, garantido que hajam poucos segmentos falsos negativo. Como exemplo supondo as seguintes atribuições para o segmento (b representando um resíduo de fita, H um resíduo de hélice e '.' um resíduo de *loop*):

```
PDB:    ...bbbbb..HHH....bbbbHHH.........HHHHbbbb.........bbb.
DSSP:   ..bbbbb...bHHHH..bbbbHHHH........HHHHHHbbb.........bbb.
União: ..bbbbbb..!HHHH..bbbbHHHH........HHHHH!bbb.........bbb.
```

Na união marcou-se os conflitos com ! e regiões de *loops* possuem menor prioridade caso outra anotação atribua uma fita ou hélice para determinado resíduo.

- `uniao_distancia.pl`
  Script criado para analisar as sequências de união de estrutura secundária e gera distribuição de distâncias em resíduo entre fitas, hélices e entre fitas com hélices.
- `extrair_estrutura.pl`
  Esse scrip foi empregado para buscar e identificar por transições entre estruturas secundárias, como exemplo transição de uma fita para uma hélice. Resíduos na vizinhança da transição são armazenados. Possui como parâmetro o número máximo de resíduos do *loop* entre a transição. Outro parâmetro é o ângulo entre as estruturas para tentar identificar transições lineares. Terceiro, é usado um parâmetro para a distância euclidiana máxima de resíduos da vizinhança até a transição. Por último o script é ajustado para buscar descontinuidades segundo um ângulo máximo formado entre os eixos centrais das estruturas secundárias. Esse último tem por objetivo que em nossos experimentos sejam buscar descontinuidades aproximadamente lineares.

- `CSM.pl`
  Gera matriz do Cutoff Scanning Matrix (CSM) [52] para cada base de dados.

Durante o desenvolvimento das atividades foi feito um desvio do plano de trabalho inicial para análise das distribuições de distância geradas a partir da base de dados. Na tabela 2 está a metodologia para cálculo de distâncias entre fitas e hélice. Para cálculo de distância entre fitas com fitas e hélices com hélices o método é semelhante, exceto que se extrai distâncias apenas de um lado da estrutura (para não ocorrer duplicações de distâncias na saída).

Um exemplo do método usado no cálculo de distâncias entre fitas e hélices:

Entrada: ...bbbbb..HHH....bbbbHHH.....HHH!bbbb.....bbb.
Saida: n 2 4 0 -1 n n n

Cada número em pares, significa a distância em resíduo de uma fita com sua hélice ascendente (primeiro numero) com sua hélice descendente. Caso não haja uma hélice ascendente ou descendente (quando o que a antecede ou precede é uma outra fita ou final de cadeia) é dado uma distância n. Seguindo o exemplo, a primeira fita está no início da cadeia então seu primeiro dígito fica com n, o segundo com 2 pois é a distancia da hélice descendente. A segunda fita tem 4 resíduos de *loop* entre a hélice ascendente e 0 para a hélice descendente. A terceira fita tem -1 de distancia da hélice ascendente

(superposição de fica com hélice no processo de união), mas não possuiu hélice descendente (o próximo seria outra fita). Por último a quarta fita não tem nem hélice ascendente nem descendente (novamente final de cadeia).

Há ainda diversas opções e configurações[1] possíveis que podem ser exploradas e configuradas utilizando os métodos das classes, como tamanho mínimo para uma fita ou hélice ser aceita.

1. Análises estatísticas em R das distribuições de distâncias
   Foi analisada a distribuição de distâncias entre fitas com hélices, fitas com fitas e hélices com hélices. Inicialmente tínhamos obtido um perfil de densidade interessante que sugeria um decaimento potencial [47] ou exponencial de distância entre fitas e hélices. Foi feito também um estudo nos gráficos de densidade acumulada, tentando aproximar possíveis parâmetros de uma distribuição potencial ou exponencial com métodos analíticos (Apêndice A e B). Foi testado ajustes com outros tipos de distribuições como Log-normal e Gumbel mas tão foram obtidos ajustes de qualidade para nenhuma dessas distribuições.

IV - Resultados obtidos e análise dos mesmos

No início do projeto foi feito um estudo da linguagem de programação Perl, estudo de arquivos pdb's, linguagem de programação R para análises estatísticas e estudo de estrutura de proteínas. Para o aprendizado sobre arquivos pdb e sua devida manipulação foi buscada e lida toda uma literatura além de participação do aluno no Curso de Verão de Bioinformática Estrutural da Universidade Federal de Minas Gerais (Apêndice C).

Durante o estudo de Perl foi decidido aplicar paradigma de orientação a objetos [48] nos scripts a serem gerados e já presentes no PDBEST. Foi também adotado uma hierarquia de diretórios para armazenamento dos arquivos gerados devido ao grande volume de dados (automaticamente gerados, se necessário, a partir do diretório do PDBEST).

Foi feito um longo processo de filtragem (tabela 1) na base de dados. Inicialmente foram removidas estruturas contendo modelos para análise futura. Em seguia foram separas as cadeias das proteínas em arquivos distintos. Foi efetivamente removido cadeias de DNA e RNA da base dados (que contendo 5 ou menos aminoácidos). Por último removemos cadeias com baixa porcentagem de fitas.

Buscamos métodos automáticos de anotar a estrutura secundária. Optamos por empregar o tradicional DSSP junto com STRIDE e de produção mais recente Kaksi. Tentamos também obter uma compilação de Votap [49] mas sem sucesso. Este está

[1] A classe Configuration.pm contida no link fornecido gerencia todas as configurações de todos estes scripts de aplicação e está bem documentada.

disponível exclusivamente como ferramenta pelo Voro3D [50] disponível apenas para Windows-MS. Priorizamos uma maior cobertura de estrutura secundaria, querendo minimizar os falsos negativos ao custo de mais falsos positivos. Ou seja, arriscamos fazer associações de SSE que podem não ser verdadeiras, apostando que estaremos "cobrindo" uma maior parte daquilo que se acredite ser SSE verdadeiras. Para isso aplicamos diferentes algoritmos de identificação de SSE, anotações automáticas (DSSP, STRIDE, Kaksi) e anotações (em algum grau) humanas (que vêm nativas no PDB). Foi feito um OR (operador binário) das diferentes associações, aceitando aquilo que pelo menos um dos métodos acreditou ser uma SSE. Em caso de associações divergirem quando a um resíduo ser hélice ou fita, foram tratados como conflitos ou superposições. Considerando o resultado desse OR, geramos algumas distribuições a partir da base de dados (figura 3). Foi também gerada uma distribuição do número de conflitos por cadeia (figura 4), representando uma pequena fração do total de resíduos.

Não foram registrados casos de um longo intervalo de uma cadeia ocorrer divergências entre dois métodos. A maior super-posição de fita por hélice (5 resíduos) foi da cadeia 1XVSA (figura 5).

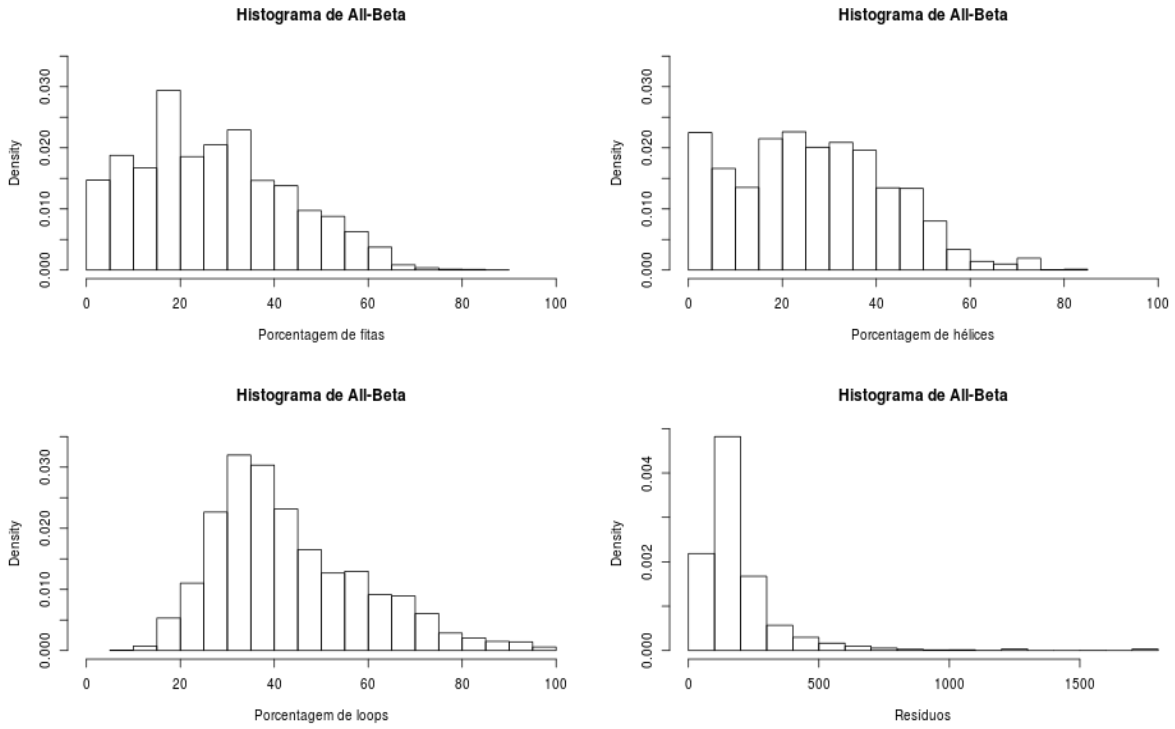


Figura 03: Distribuições geradas na base de dados filtrada. Distribuição de porcentagem de fitas, porcentagem de hélices, porcentagem de resíduos em *loop* e distribuição de número de resíduos. Pode-se notar que após a filtragem temos uma distribuição sem cadeias com 100% de fita nem hélices e nem cadeias predominantemente compostas por *loops*. A análise de que se algumas dessas distribuições sigam uma função de probabilidade talvez mereça uma atenção futura (como qui-quadrado para a distribuição de resíduos, e talvez normalmente distribuído para a porcentagem de resíduos em *loop*).

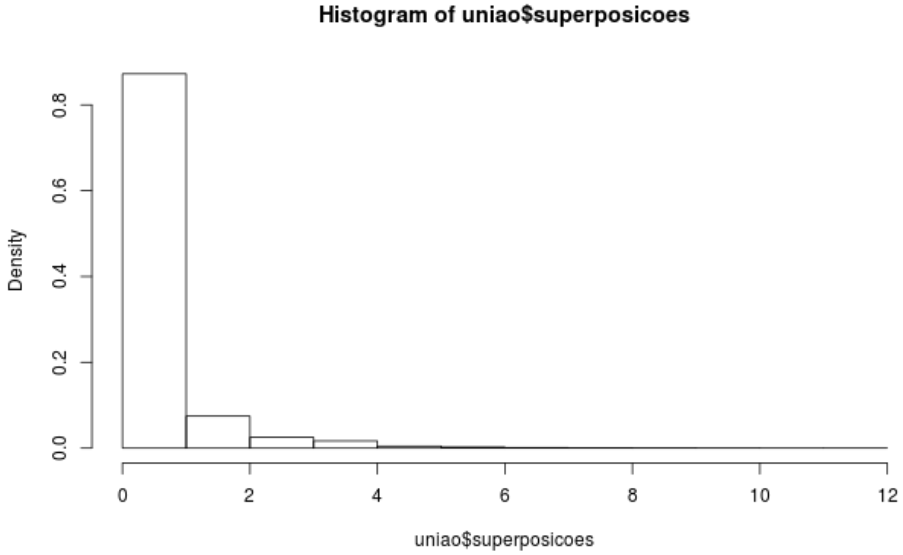


Figura 4: Distribuição de conflitos. Podemos notar que a maior parte das cadeias não possui associações divergentes. Foi gerado um total de 3580 associações divergentes o que corresponde a apenas 0,31% de todos os resíduos presentes nas cadeias da base de dados.

Figura 5: Cadeia em que ocorre o maior conflito (5 resíduos). STRIDE associa uma continuação de uma fita antes do encontro a uma hélice, enquanto os anotadores do PDB associam interrupção para uma hélice. Pela imagem nota-se que a hélice presente entre as duas fitas acaba apresentando uma forma estranha, bem distorcida.

Utilizando as sequências SSE geradas pela união das atribuições de anotadores automáticos e o nativo do PDB, foi gerada uma distribuição das distâncias entre fitas, hélices e entre fitas com hélices (Figura 6). Inicialmente o tamanho de um conflito foi tratado como uma distância negativa entre fitas e hélices. Posteriormente essas distâncias negativas foram convertidas para 0 para ajuste de uma função de densidade para distribuição discreta.

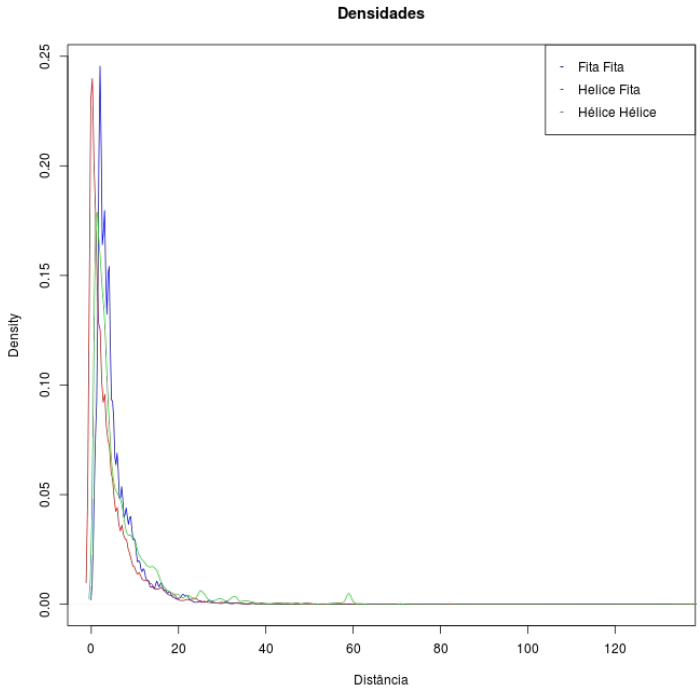


Figura 6: Comparativo gráfico das distribuições de distâncias. As 3 distribuições parecem se ajustar bem entre si.

Tentou-se ajustes para uma função de probabilidade para a distribuição de distância entre fita e hélice. Inicialmente tentamos um ajuste a uma lei de potência (foi preciso remover as distâncias iguais a 0 pois a função potencial não está definida nesse ponto), obtendo bons resultados para $x \geq 15$, o que no entanto corresponde a apenas aproximadamente 5% de toda a distribuição, indicando que a distribuição como um todo não se ajusta a uma lei de potência. Utilizamos scripts R fornecidos por Clauset [47] que permite realizar esse ajuste e gerar distribuições artificiais. Empregamos as fórmulas obtidas pelos cálculos descrito no Apêndice A, para chegar em parâmetros precisos usados para gerar distribuições artificiais que se aproximam bem da distribuição de distância entre fitas e hélices (Figura 7) para valores de x maiores ou iguais a 15. Foi estimado um ajuste de alpha de 3.57 com um xmin de 15.

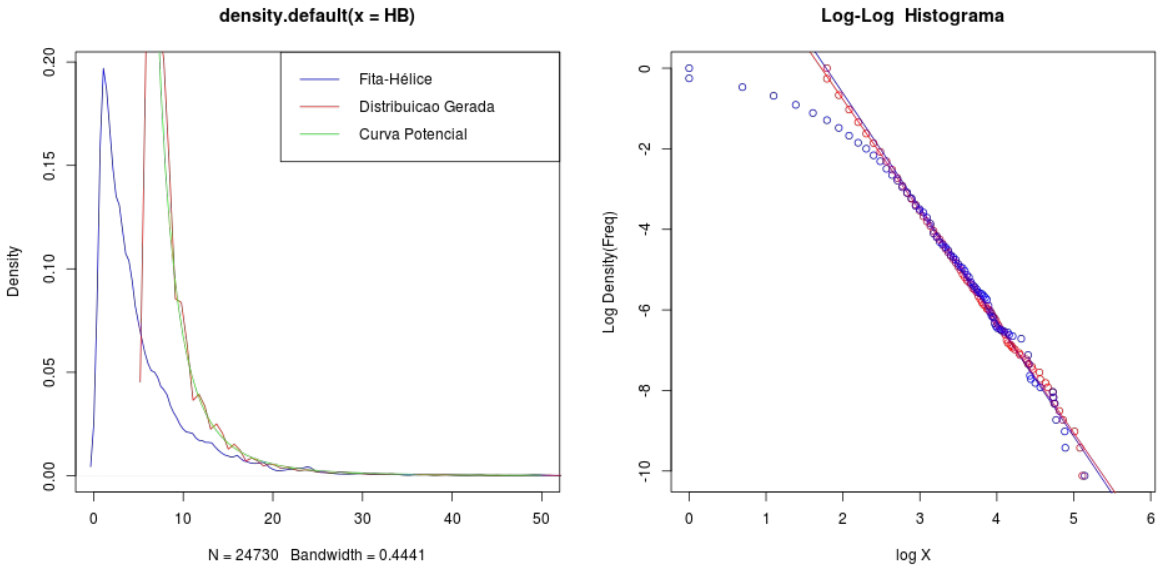


Figura 7: No gráfico da esquerda temos a distribuição de distância entre fitas e hélices (azul), a distribuição gerada a partir dos parâmetros calculados com dados extraídos da regressão linear do gráfico de densidade acumulada em loglog (vermelho), curva potencial pura correspondente à distribuição gerada (verde). Vemos pelo segundo gráfico que a distribuição de distâncias só segue uma reta a partir de um valor estimado como sendo 15 (a regressão foi realizada somente para valores acima desse ponto). A distribuição gerada gera uma reta de regressão muito próximo da distribuição real e pelo primeiro gráfico vemos que isso ocorre a partir de uma distância entre 15 e 20.

Também foram testado outros possíveis ajustes. Graficamente conseguimos um ajuste para toda a distribuição empregando um ajuste exponencial (Figura 8) utilizando a biblioteca igraph [51] para a linguagem R. Foi estimado um rate de 0.17, sendo preciso ainda avaliar a qualidade do ajuste, no entanto não foi obtido um bom valor para *goodness of fit*.

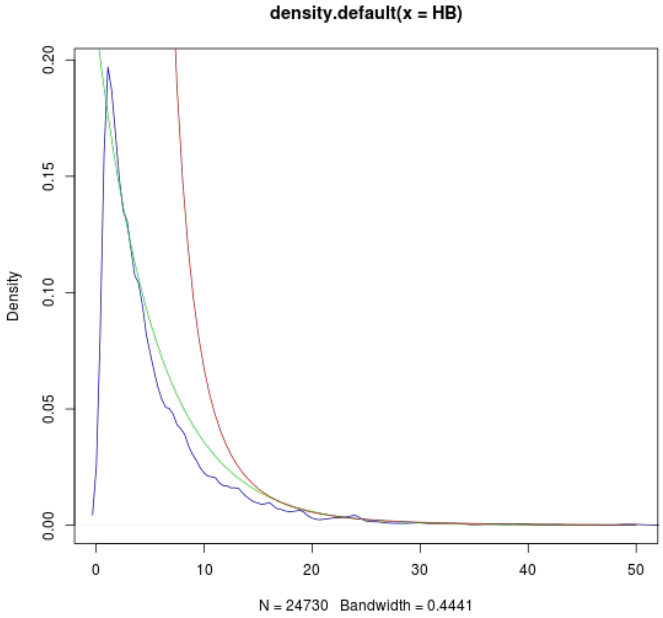


Figura 8: Comparação entre ajuste exponencial (verde) e potencial (vermelho). A curva exponencial parece se ajustar melhor que a curva potencial desde o início da distribuição.

Iniciando a etapa de mineração de dados foi criado o script para identificar regiões de transição. Foram geradas 4 bases de dados com arquivos descrevendo regiões de transição de fita para hélice, hélice para fita, hélice para hélice e fita para fita usando as 5873 cadeias filtradas. Foi usado um parâmetro *loop_máximo* = 10 resíduos para controlar o tamanho em resíduo da descontinuidade, por exemplo entre duas fitas. Em seguida para esses resíduos foi selecionado na estrutura tridimensional todos os aminoácidos que estavam a uma distância máxima de 30 Ångstroms. Somente os aminoácidos da região e os vizinhos foram salvos nos arquivos que compõem as bases de dados.

Quantidade de regiões armazenadas nas bases:

- hélice para fita: 5919 regiões
- fita para hélice: 6427 regiões
- fita para fita: 5142 regiões
- hélice para hélice: 3833 regiões

Com as bases de dados das transições de estrutura secundária foram geradas matrizes do CSM [52]. O perfil das distribuições de densidade (figura 9) se mostraram poucos sensíveis à variação do tamanho do *loop* máximo. É possível observar que as distribuições médias (figura 10) entre as 4 bases de dados se mostraram bem próximas divergindo possivelmente devido à distribuição do tamanho das cadeias (figura 11) para cada uma das bases de dados.

Paramos na etapa de aplicação do SVD [53-55] para diminuir ruído e redução de dimensões na matriz CSM gerada. Pretendemos encontrar o melhor ponto de corte nos

valores singulares (figura 12) gerados para obter maior precisão na informação contida nos dados iniciais.

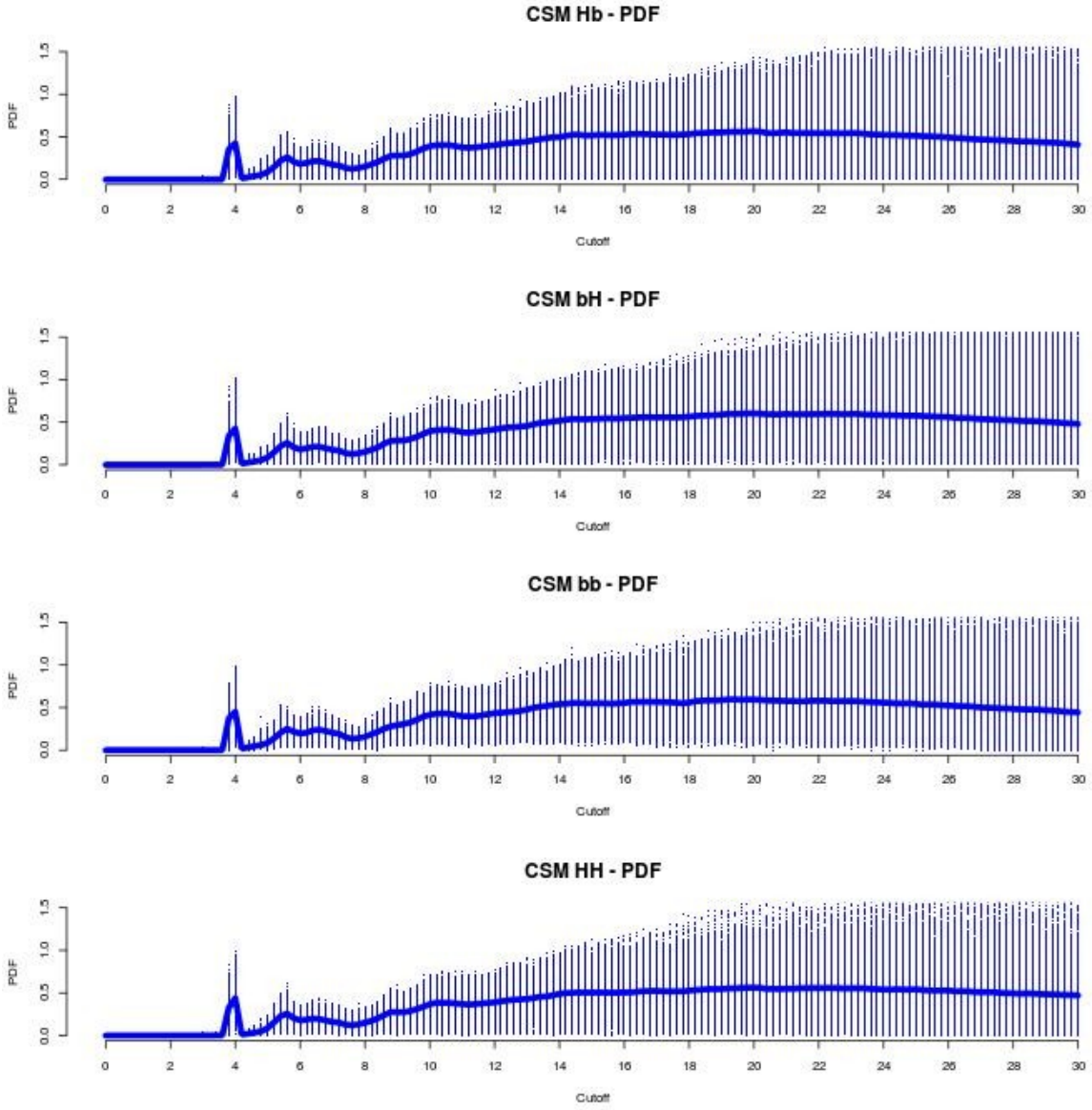


Figura 9: Distribuição do CSM para as bases de dados de transição fita para hélice (bH), hélice para fita (Hb), fita para fita (bb) e hélice para hélice (HH). Bases de dados geradas para um *loop* máximo de 10 resíduos e distância de carbonos alfas vizinhos à transição de 30 Ångstrons. As frequências do CSM foram normalizadas segundo o tamanho da cadeia que se encontra a transição. Ao aplicar o CSM foi empregado um cutoff mínimo de 0,0Å e máximo 30,0Å com um passo de 0,2Å.

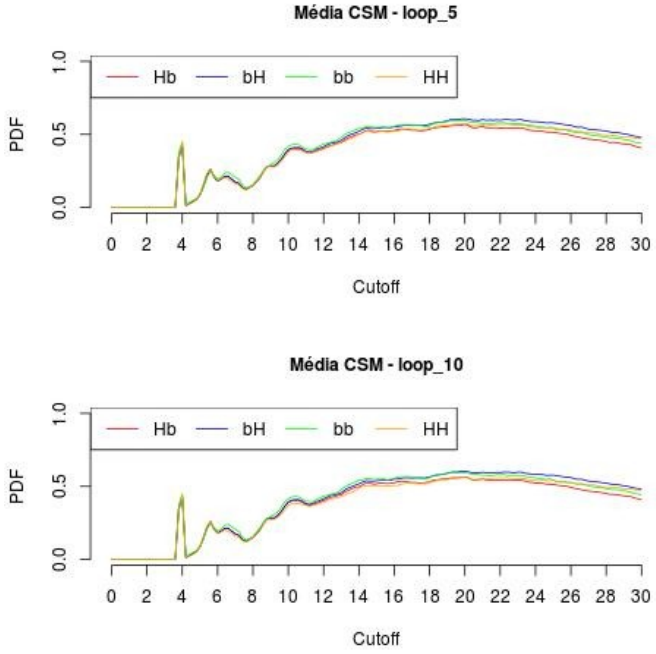


Figura 10: Comparação das distribuições médias do CSM para bases com transições com *loop* máximo de 5 e 10 resíduos. Vemos pouca diferença antes do *cutoff* de 12 Ångstrons. Além disso o perfil das distribuições com *loops* máximos diferentes é bem parecido, afetando apenas a quantidade de transições nas bases de dados.

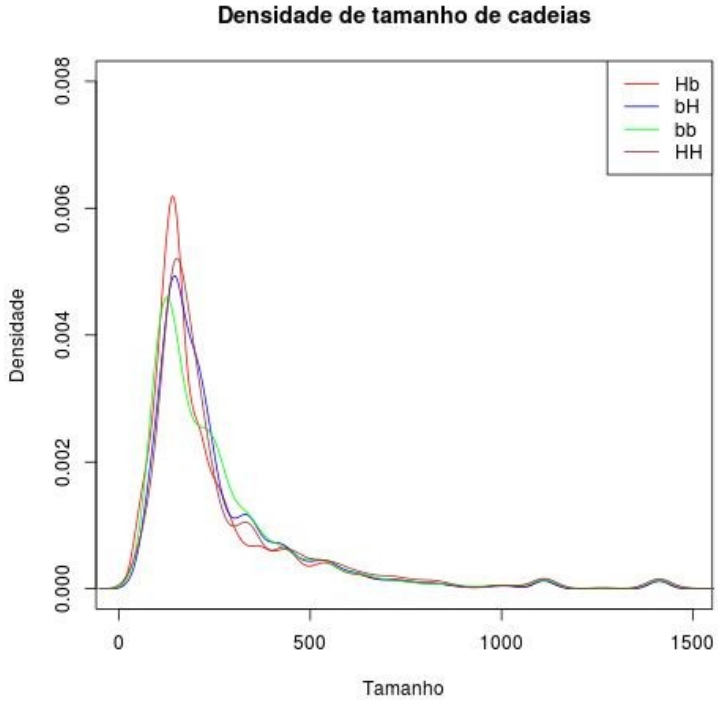


Figura 11: Distribuição do tamanho das cadeias para as bases de dados de transições usadas para normalizar a distribuição do CSM.

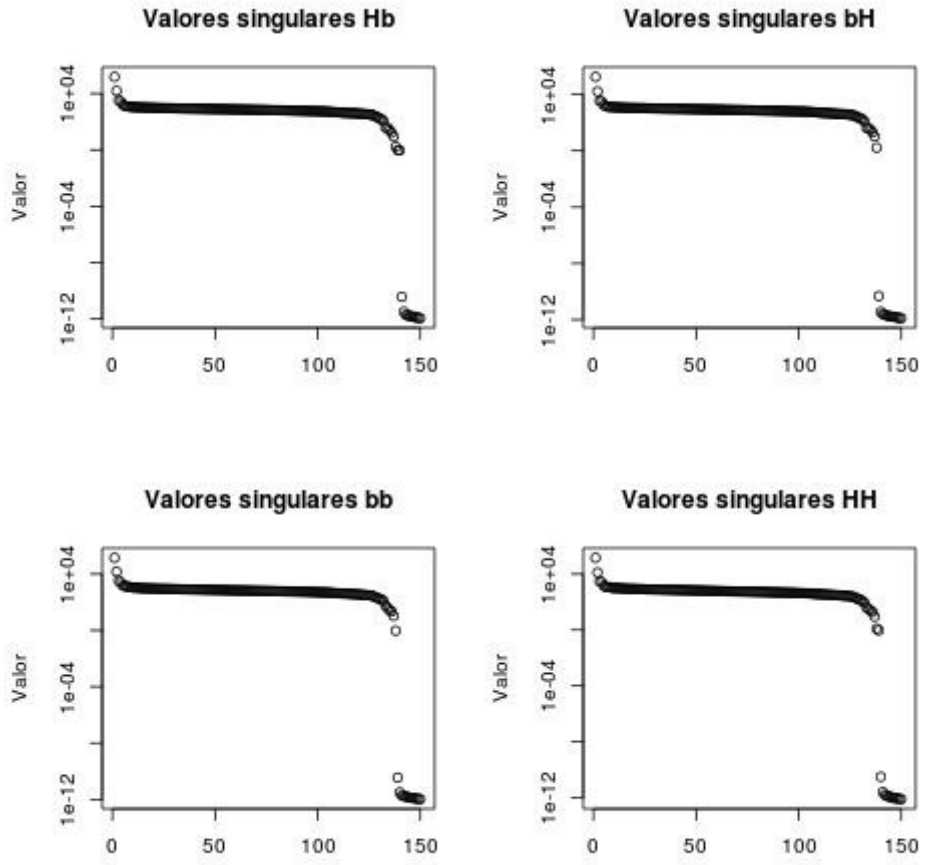


FIgura 12: Valores singulares para a matriz do CSM para as 4 bases de dados. Há 151 valores singulares referentes aos intervalos de *cutoff*. Selecionando uma quantidade suficiente de valores singulares é possível obter uma matriz conservando as propriedades dos dados iniciais com menos dimensões e com redução de ruídos. O eixo Y possui uma escala logarítmica.

V - Conclusões e metas futuras

Durante o desenvolvimento foi necessário buscar uma grande gama de conhecimento para o andamento no projeto. A vasta quantidade de dados disponíveis gratuitamente pela internet (PDB) tornou fácil a montagem da base de dados local. Conseguimos realizar uma boa filtragem utilizando scripts perl enquanto observada a qualidade da base de dados utilizando análises estatísticas em R.

Durante a vigência do projeto não foi possível realizar todas as atividades propostas devido a sua complexidade prática e teoria, assim como o grande número de etapas que foram realizadas. Além disso, durante o projeto desviamos do foco inicial para analisar os interessantes resultados obtidos pelas distribuições geradas. Em destaque a distribuição de

distâncias entre fitas e hélices em que foi possível um bom ajuste parcial para uma distribuição potencial.

Pretendemos modificar a abordagem inicial de analisar cadeias e passar a analisar domínios, tal como faz o SCOP[20]. Isso será feito para todos os tipos de SSE, mas tem um efeito especial em *all-beta*. Deve-se ao fato de ao separarmos as cadeias de proteínas *all-beta* nem todas as cadeias apresentam fitas, sendo necessário remover uma grande quantidade de cadeias de DNA e cadeias com pequena incidência de fitas.

Com as bases de transições geradas pretendemos comparar as assinaturas de diferentes tipos de transições. Pretendemos também investigar a vizinhança das regiões de transições de fitas para hélices para testar a hipótese de que as descontinuidades se relacionam ao arranjo tridimensional de aminoácidos. Antes de iniciar o processo de classificação utilizando os dados do CSM pretendemos aplicar a redução de dimensionalidade do SVD para redução de ruídos e poupar desempenho computacional.

Diante dos resultados obtidos pretendemos publica-los e é de interesse posteriormente disponibilizar as ferramentas do PDBEST via internet.

Pretendemos também apresentar esses resultados preliminares no X-Meeting 2013 (http://evento2013.x-meeting.com/), que este ano será uma *International Conference* envolvendo a *Brazilian Association for Bioinformatics and Computational Biology* (AB3C) e o *Brazilian Symposium on Bioinformatics* (BSB), em Recife, Pernambuco, de 03 a 07 de Novembro de 2013.

VI - Metodologia e Plano de Trabalho

O projeto está divido em três fases, subdividida em etapas, e faz uso de uma metodologia de pesquisa e desenvolvimento por prototipagem incremental:

FASE I	**Montagem da Base de Dados**
~~Etapa 01~~	~~Levantamento Bibliográfico~~
~~Etapa 02~~	~~Construir uma base de dados, a partir do PDB – Protein Data Bank, contendo as cadeias toda-beta com transições fita-hélice de interesse~~
~~Etapa 03~~	~~Construir scripts PERLs que auxiliem no processo de filtragem das cadeias selecionadas na etapa anterior~~
~~Etapa 04~~	~~Aplicar algumas métricas estatísticas para verificar a qualidade da base de dados.~~
FASE II	**Implementações e Testes**
~~Etapa 05~~	~~Em PERL, montar scripts capazes de identificar e caracterizar as regiões de transição de fitas para hélices.~~

Etapa 06	Aplicar técnicas de mineração de dados capazes de levantar padrões envolvendo parâmetros diversos da sequência e estrutura nas regiões de transição fita/hélice. Verificar se esses padrões podem sugerir um nova metodologia para identificação e caracterização de estruturas secundárias em proteínas.
FASE III	**Avaliação**
Etapa 07	Apresentar os resultados obtidos até então no X-Meeting, a principal conferência internacional da área no Brasil, por volta de outubro/novembro de 2012. Participar da Jornada de Iniciação Científica, Tecnológica e Inovação, da UNIFEI.
~~Etapa 08~~	~~Avaliar os resultados. Produzir relatórios.~~

VII - Cronograma de Atividades

Fase - Etapa	1	2	3	4	5	6	7	8	9	10	11	12
I - 01												
I - 02												
I - 03												
I - 04												
II - 05												
II - 06												
III - 07												
III - 08												

VI - Referências Bibliográficas

[1]BUENO, F. S. Grande dicionário etimológico-prosódico da língua portuguesa - vol. IV. São Paulo: Lisa, 1988, p. 1826.

[2]LEHNINGER, A. L., NELSON, D. L., COX, M. M.. Principles of biochemistry. 2.ed. New York : Worth Publisher, 1993

[3]GREIGHTON, T. E. Proteins: structures and molecular principles. New York : W. H. Freeman and Co, 1983.

[4]PAULING, L The Structure of Protein Molecules. Scientific American, jul 1954.

[5]KENDREW, J.C., et al. A three-dimensional moldel of the myoglobin molecule obtained by x-ray analysis. Nature 181: 662-666, 1958.

[6]PERUTZ, M. F., KENDREW, J.C., WATSON, H.C., J. Mol. Biol., v.13, p.669, 1965.

[7]PRIVALOV, P. L., KHECHINASHVILI, J. A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. J. of Mol. Biol., N.86, 1974

[8]SANTORO, M. M., BOLEN, D.N. A test of the linear extrapolation of unfolding free energy changes over an extended denaturant concentration range. Biochemistry, n.31, p.4901-4907, 1992.

[9]RAMOS, C. H., KAY, M.S., BALDWIN, R.L.. Putative interhelix íon pairs involved in the stability of myoglobin. Biochemistry, v.30, n38, p.9783-90, jul.1999

[10]LEVINTHAL, C. How to fold graciously ? Proceedings of Mössbauer Spectroscopy in Biological Systems Meeting, p22-24, 1967.

[11]ANFINSEN, C.B. Principles that govern the folding of protein chains. Science, v. 181, p.223-230, 1973

[12]RICHARDS, F. M. The Protein folding problem. Scientific American, p.34-41, jan. 1991

[13] CREIGHTON, T. E. The protein folding problem. Science, v. 240, p. 267-344, 1988.

[14] PERUTZ, M.F. BIOGRAPHY. In: Nobel Lectures, Chemistry 1942-1962, Amsterdam: Elsevier publishing company, 1964.

[15] KENDREW,J.C. BIOGRAPHY. In: Nobel Lectures, Chemistry 1942-1962, Amsterdam: Elsevier publishing company, 1964.

[16] WÜTHRICH, K. Autobiography. In: Les Prix Nobel. The Nobel Prizes 2002, Stockholm: Editor Tore Frängsmyr, 2003.

[17] WHITE, S. H.Translocons, thermodynamics, and the folding of membrane proteins. FEBS Letters, v. 555, p. 116-121, 2003.

[18] DOMAN, T. N. et al. Molecular Docking and High-Throughput Screening for Novel Inhibitors of Protein Tyrosine Phosphatase-1B. Journal of Medical Chemistry, v. 45, p. 2213-2221, 2002.

[19] BERNSTEIN, F. C.; KOETZLE, T. F; WILLIAMS, G. J.; MEYER JR, E. F.; BRICE, M. D.; RODGER, J. R.; KENNARD, O.; SHIMANOUCHI, T.; TASUMI, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. Journal of Molecular Biology, v. 112, n. 3, p. 535–542, 1977.

[20] MURZIN, A. G.; BRENNER, S. E.; HUBBARD, T.; CHOTHIA, C. SCOP: a structural classification of proteins database for investigation of sequences and structures. Journal of Molecular Biology, v. 247, p. 536-540, 1995.

[21] PEARL, F. M.; BENNETT, C. F.; BRAY, J. E.; HARRISON, A. P.; MARTIN, N.; SHEPHERD, A.; SILLITOE, I.; THORTON, J.; ORENGO, C. A. The CATH database: an extended protein family resource for structural and functional genomics. Nucleic Acids Researchs, v. 31, p. 452-455, 2003.

[22] PDB - Protein Data Bank. Disponível em: http://www.pdb.orb. Acessado em: novembro de 2011.

[23] KABSCH, W.; SANDER, C. Dictionary of protein secondary structure: pattern recognition hydrogen-bonded and geometrical features. Biopolymers, v. 22, p. 2577-2637, 1983.

[24] FRISHMAN, D.; ARGOS, P. Knowledge-based protein secondary structure assignment. Proteins: Structure, Function and Genetics, v. 23, p. 566-579, 1995.

[25] DILL, K. A. et al. The protein folding problem: when will it be solved? Current Opinion in Structural Biology. v. 17, p. 342–346, 2007.

[26] KARPLUS, M. WEAVER, D. L. Protein folding dynamics: the diffusion-collision model and experimental data. Protein Science, v.3, p.650–68, 1994.

[27] DAGGETT, V.; FERSHT, A. R. Is there a unifying mechanism for protein folding? TRENDS in Biochemical Sciences, v. 28, n. 1, p. 18-26, 2003.

[28] DILL, K. A. Polymer principles and protein folding. Protein Science, v. 8, p. 1166-1180.1999.

[29] CREIGHTON, T. E. How important is the molten globule for correct protein folding? Trends in Biochemistry - TIBS, v. 22, p. 6-10, 1997.

[30] BALDWIN, R. L.; ROSE, G. D. Is protein folding hierarchic? I. Local structure and peptide folding. TIBS,v. 24, p. 26-33, 1999.

[31]RICHARDS, F. M. Protein stability: still an unsolved problem. Cell Molecular Life Science - CMLS, v. 53,
p. 790-802, 1997.

[32] HONIG, B. Protein folding: from the Levinthal paradox to structure prediction. Journal of Molecular Biology, v. 293, p. 283-293, 1999.

[33] MUNOZ, V., SERRANO, L. Elucidating the folding problem of helical peptides using empirical parameters. Journal of Molecular Biology. v. 245, p. 275-96, 1995.

[34] ROSE, G. D., SRINIVASAN, R. Ab initio prediction of protein structure using LINUS. Proteins: Structure, Function, and Bioinformatics. v. 47, n. 4, p. 489–495, 2002

[35] VALENCIA, A. Protein refinement: a new challenge for CASP in its 10th anniversary. Bioinformatics, v.21, n. 3, p. 277, 2005.

[36]Turcotte, M.; Muggleton, S. H. & Sternberg, M. J. E. Automated discovery of structural signatures of protein fold and function. Journal of Molecular Biology, n. 306, v. 3, p. 591-605, 2001.

[37]Kersting, K.; Raiko, T.; Kramer, S. & De Raedt, L. (2003). Towards discovering structural signatures of protein folds based on logical hidden markov models. Em Proceedings of the Pacific Symposium on Biocomputing, pp. 192-203, 2003.

[38] MOTTALIB, M.A., et al Protein Secondary Structure Prediction using Feed-Forward Neural Network. ISSUE, V. 01, 2005.

[39] DOBSON, P.D, DOIG, A.J, Predicting Enzyme Class From Protein Structure Without Alignments. J. Mol. Biol. v. 345, p. 187–199, 2005.

[40] [31] BALDWIN, R. L.; ROSE, G. D.Is protein folding hierarchic? II. Folding intermediates and transition states. TIBS,v. 24, p. 77-83, 1999.

[41] MARTIN, Juliette. et al. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Structural Biology, 5:17, 2005.

[42] Pires, D. E. V.; Da Silveira, C. H.; Meira Junior, W.; Santoro, M. M. PDBEST - PDB Enhanced Structures Toolkit. In: X-Meeting 2007 - 3rd Annual Conference of the AB3C, 2007, São Paulo. Poster Abstract. Chosen as the best work in the databases category.

[43] BRANDEN, Carl. TOOZE, John. Introduction to protein structure 2nd ed. Garland Publishing: New York, NY. 1999.

[44] KABSCH, Wolfgang. SANDER, CHRISTIAN. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. Biopolymers, Vol. 22, 2577-2637, 1983.

[45] Frishman D, Argos P. (1995). Knowledge-based protein secondary structure assignment. *Proteins* 23(4):566-79. doi:10.1002/prot.340230412 PMID 8749853.

[46] IHAKA, Ross. GENTLEMAN, Robert. A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics. Vol. 5, n. 3, 299-314, 1996.

[47] CLAUSET, Aaron. SHALIZI, Cosma Rohilla. NEWMAN, M. E. J. Power-Law distributions in empirical data. ArXiv: 0706.1062v2, 2009.

[48] CONWAY, Damian. Object Oriented Perl. Manning Publications, 1999.

[49] POUPON,A. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. Curr. Opin. Struct. Biol., 14, 233–241, 2004.

[50] DUPUIS, Franck. et al. Voro3D: 3D Voronoi tessellations applied to protein structures. Bioinformatics, vol. 21 no. 8, 1715-1716, 2005.

[51] CSÁRDI, Gábor. NEPUSZ, Tamás. The igraph software package for complex network research. InterJournal Complex Systems, 1695, 2006.

[52] Pires, D. E. V.; Da Silveira, C. H.; Meira Junior, W.; Santoro, M. M; Melo-Minardi, Raquel; Santos, Marcos A. Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance pattern. In: X-Meeting 2010 - 6rd Annual Conference of the AB3C, 2007, Ouro Preto.

[53] Eldén L: Matrix Methods in Data Mining and Pattern Recognition (Fundamentals of Algorithms). Society for Industrial and Applied Mathematics 2007.

[54] Eldén L: Numerical linear algebra in data mining. Acta Numerica 2006, 15:327-384.

[55] Berry MW, Dumais ST, O'Brien GW: Using linear algebra for intelligent information retrieval. SIAM review 1995, 37(4):573-595.

APÊNDICE A

Método usado para estimar parâmetros de uma *power-law* que se ajusta graficamente aos dados. Primeiro mostramos que ao aplicar Log em ambos os eixos da distribuição de densidade acumulada (a distribuição de densidade é muito sensível a pequenas variações, então optou-se por usar a distribuição acumulada) espera-se que a distribuição siga uma reta. Em seguida utilizando dos parâmetros de regressão estimados como intercepto e inclinação é possível mostrar qual deve ser o menor valor (xmin) e qual o expoente para que uma distribuição artificial gerada se sobreponha bem aos dados reais. Também é calculada a constante da lei de potencial como função do expoente e do xmin da distribuição.

A distribuição sendo potencial de forma

$$f(x) = \frac{c}{x^a} \qquad \text{(Equação 1)}$$

Integrando de uma variável x até infinito, em um intervalo [b, Inf) obtêm-se a função acumulada. Sendo b o menor valor dessa distribuição (vamos chamar de xmin').

$F(x) = \int_x^\infty \frac{c}{t^a} dt = \frac{c}{1-a} \lim_{n\to\infty} n^{1-a} - x^{1-a}$, resolvendo o limite (a > 1)

$$F(x) = \frac{c}{a-1} x^{1-a}, b \leq x < \infty \qquad \text{(Equação 2)}$$

Aplicando Log em ambos os lados da função acumulada temos a equação da reta de um gráfico em loglog da função acumulada.

$$\ln F(x) = (1-a)\ln x + \ln \frac{c}{a-1} \qquad \text{(Equação 3)}$$

Sendo b o menor valor da distribuição, nosso xmin' (não necessariamente corresponde ao ponto inicial da distribuição em que ela passa a seguir uma lei de potência). Temos ainda a função que gera aleatoriamente uma distribuição que segue uma lei de potencial em que o parâmetro b, é o menor valor que essa função ira gerar, novamente 'xmin'.

$$plrand = \frac{b}{(1-x)^{\frac{1}{a-1}}} \qquad \text{(Equação 4)[47]}$$

Querendo encontrar um valor de b que forneça um determinado intercepto a partir da equação da reta.

Como $b \leq x < \infty$, então a função acumulada em b é 1 (pois é uma função de probabilidade), ou seja acumula toda a distribuição.

$F(b) = 1$

$\ln F(b) = 0 = (1-a)\ln b + \ln \frac{c}{a-1}$

$(a-1)\ln b = \ln \frac{c}{a-1}$

$\ln b^{a-1} = \ln c - \ln(a-1)$

$\ln c = \ln[b^{a-1}(a-1)]$

$$c = b^{a-1}(a-1) \qquad \text{(Equação 5)}$$

a constante está relacionada a xmin e o valor de alpha

Para um determinado intercepto, da equação da reta $\ln \frac{c}{a-1} = d$ queremos o valor de b que forneça uma distribuição de intercepto d.

Manipulando a equação (5):
$\frac{c}{a-1} = b^{a-1}$, e aplicando log
$\ln \frac{c}{a-1} = \ln b^{a-1}$, o lado esquerdo da equação se resume a 'd', logo isolando b:
$\frac{d}{a-1} = \ln b$

$$b = e^{\frac{d}{a-1}} \quad \text{(Equação 6)}$$

Para dado valor de alpha, a equação da reta fornece uma inclinação igual a:

$$m = 1 - a \quad \text{(Equação 7)}$$

APÊNDICE B

Método usado para estimar parâmetros de uma exponencial que se ajusta graficamente aos dados. Primeiro mostramos que ao aplicar Log no eixo Y da distribuição de densidade acumulada (a distribuição de densidade é muito sensível a pequenas variações, então optou-se por usar a distribuição acumulada) espera-se que a distribuição siga uma reta. Em seguida utilizando dos parâmetros de regressão estimados como intercepto e inclinação é possível mostrar qual deve ser o menor valor (xmin) e qual o expoente para que uma distribuição artificial gerada se sobreponha bem aos dados reais. Como é aplicado Log apenas no eixo Y, a distribuição é muito sensível ao valor máximo da distribuição, gerando uma reta ruim. No entanto o procedimento descrito ilustra o cálculo da constate da curva exponencial.

Tentando um ajuste para uma função exponencial do tipo:

$$f(x) = ce^{-rx} \quad \text{(Equação 8)}$$

Integrando de x até infinito temos a função acumulada (novamente para evitar oscilações ao aplicar log).

$$F(x) = \int_x^\infty ce^{-rt}dt = c \lim_{n\to\infty} \frac{e^{-rn}}{-r} - e^{\frac{-rx}{-r}}$$

$$F(x) = c\frac{e^{-rx}}{r} \quad \text{(Equação 9)}$$

Aplicando log na equação (2)

$$\ln F(x) = \ln c - \ln r - rx \quad \text{(Equação 10)}$$

Podemos encontrar essa constante com base na mesma lógica usada para encontrar a constante da lei de potência (Apêndice A). Partindo como base que a função acumulada de xmin até infinito é 1:

$$F(x_{min}) = F(b) = 1$$
$$\ln F(b) = \ln c - lnr - rb$$
$$0 = \ln c - \ln r - rb$$
$$\ln c = \ln r + rb$$
$$c = e^{\ln r} e^{rb}$$
$$c = re^{rb}$$

Com base nisso, a função se resume a:

$$f(x) = re^{rx_{min}} e^{-rx} \qquad \text{(Equação 11)}$$

Para um xmin igual a 0 a constante se resume somente ao r (rate).

APÊNDICE C

CERTIFICADO

Ministério da Educação
Universidade Federal de Minas Gerais
Pró-Reitoria de Extensão
Atualização

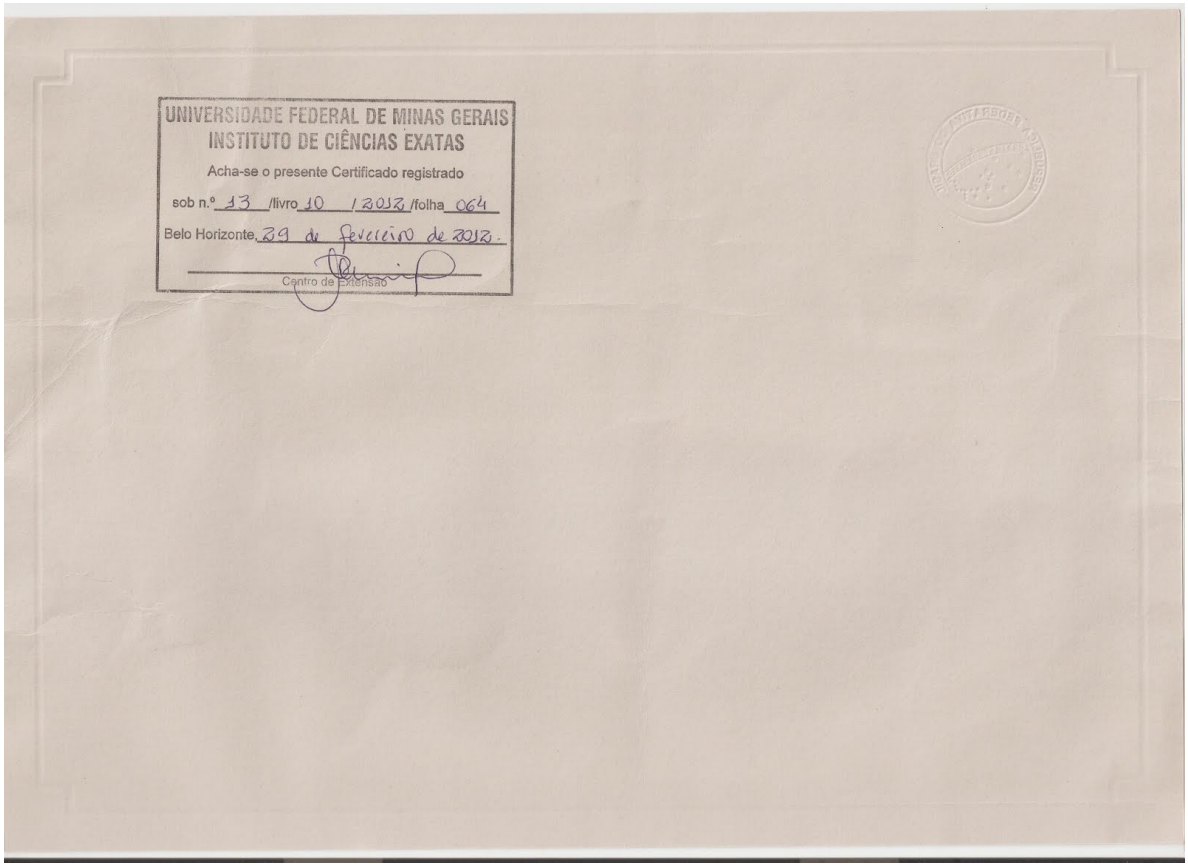
Certificamos que ***Mateus Venturyne Xavier Ferreira*** , participou como aluno do Curso de Verão em ***"Bioinformática Estrutural,"*** promovido pelo Departamento de Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, realizado no período de 06-02-2012 a 10-02-2012, com uma carga horária de 40 horas-aula.

Belo Horizonte, 23 de fevereiro de 2012.

Prof. Roberto da Silva Bigonha
Chefe do Departamento de Ciência da Computação
ICEx – UFMG

Prof. Antônio Otávio Fernandes
Diretor do Instituto de Ciências Exatas
UFMG

Prof. Wagner José Corradi Barbosa
Coordenador do CENEX - ICEx


UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
Acha-se o presente Certificado registrado
sob n.º 13 /livro 10 /2012 /folha 064
Belo Horizonte, 29 de fevereiro de 2012.
Centro de Extensão