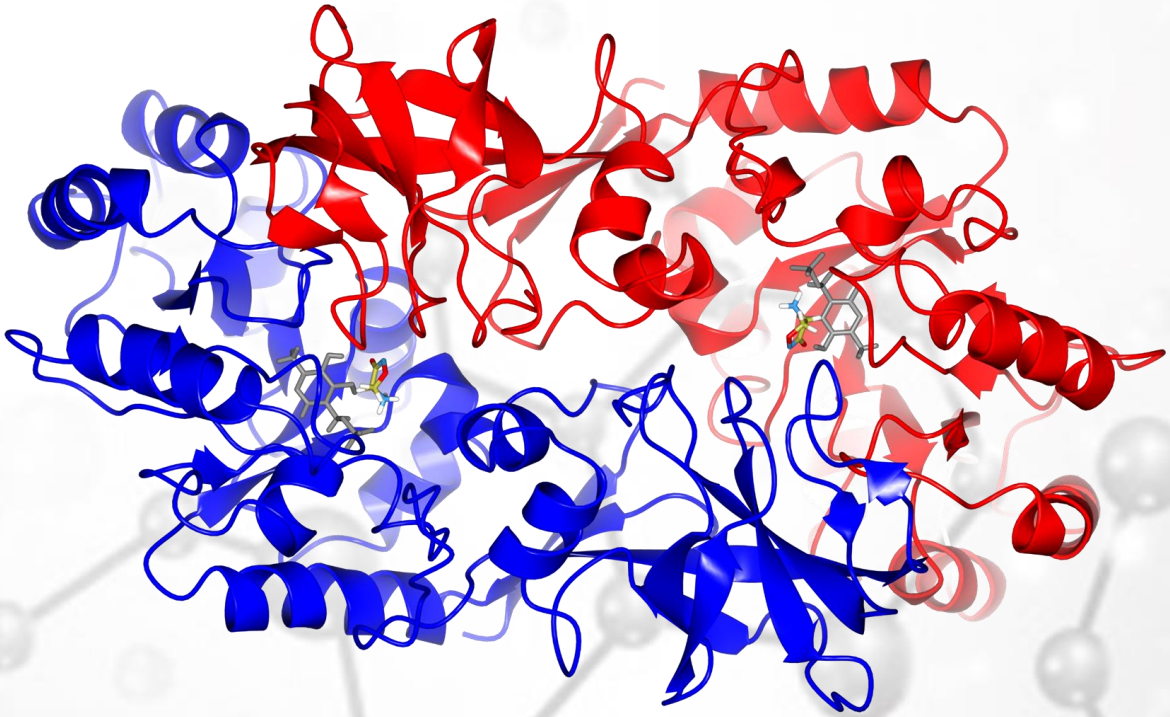
UNIVERSIDADE ESTADUAL DE MARINGÁ
Programa de Pós-Graduação em Ciências Biológicas
Área de Concentração: Biologia Celular e Molecular

VI Curso de Inverno em Biologia Celular e Molecular do PBC

25 a 29 de julho de 2016



Ferramentas de bioinformática na caracterização de alvos de medicamentos

MsCª. Arethusa Lobo Pimentel
Paulo Sérgio Alves Bueno

Prof. Dr. Flávio Augusto Vicente Seixas
Orientador

Apresentação

Neste minicurso serão apresentadas algumas ferramentas de bioinformática aplicadas na análise de proteínas, para fins de caracterização dos diferentes níveis de estrutura, iniciando com a sequência de aminoácidos e terminando na determinação da estrutura tridimensional. Esta caracterização tem uma importância biotecnológica na identificação de potenciais alvos de fármacos que podem ser utilizados em estudos "*in silico*" de varredura virtual para identificação de candidatos a medicamentos.

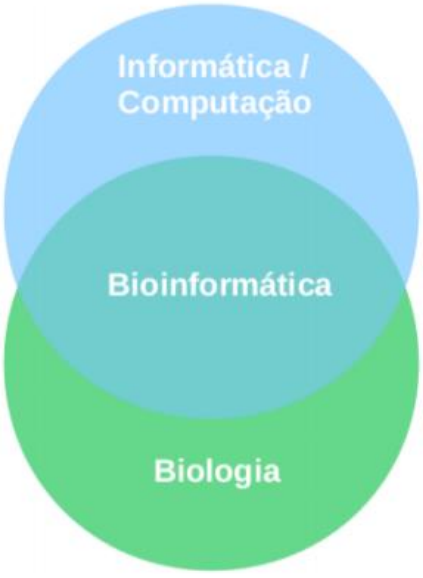
Sumário

1 O que é Bioinformática?

"Pesquisa, desenvolvimento ou aplicação de ferramentas computacionais e abordagens para expansão do uso de dados biológicos, médicos, comportamentais ou de saúde, incluindo a aquisição, armazenamento, organização, arquivamento, análise e visualização desses dados (NCBI, 2001)."

A bioinformática é a utilização de métodos computacionais, matemáticos e estatísticos para analisar dados biológicos, bioquímicos e biofísicos. Este é um campo de estudo relativamente recente, que evolui rapidamente e além disso, possui uma ampla definição devido ao vasto campo de estudo e interdisciplinaridade, dependendo da área do conhecimento ou objetivos a que se destina sua utilização. Podemos considerar a bioinformática como uma linha de pesquisa que envolve aspectos multidisciplinares e que surgiu a partir do momento em que se iniciou a utilização de ferramentas computacionais para a análise de dados genéticos, bioquímicos e de biologia molecular. Também pode ser definida como uma ciência e tecnologia de aprendizagem, gestão e processamento de informação biológica. A bioinformática é muitas vezes focada na obtenção e orientação de dados biológicos, na organização destas informações em bases de dados, no desenvolvimento de métodos para obtenção de informações úteis e a partir de tais bases de dados, a elaboração de métodos para a integração de informações relacionadas a partir de diferentes fontes. Constantemente bases de dados de computador e algoritmos são desenvolvidos para acelerar e reforçar a investigação biológica (Thampi, 2009).

A bioinformática envolve a união de diversas linhas de conhecimento – a ciência da computação, a engenharia de softwares, a matemática, a estatística e a biologia molecular – e tem como finalidade principal desvendar a grande quantidade de dados que vem sendo obtida através do sequenciamento de DNA e de proteínas. No estudo de genomas completos, a informática é imprescindível e a biologia molecular moderna não estaria tão avançada hoje, não fossem os recursos computacionais existentes. Como exemplo, as ferramentas de bioinformática são fundamentais para a genômica comparativa, utilizando conhecimentos de modelos de organismos (não humanos) para se obter informações sobre a função e as estruturas de



genes e proteínas, causas de doenças e os mecanismos da vida. Os biólogos evolucionistas utilizam a bioinformática para estudar os mecanismos da evolução através da exploração da homologia de genes ortólogos e proteínas, enquanto farmacologistas exploram os benefícios e perigos das drogas, utilizando informações sobre vias de transdução de sinais bioquímicos e biólogos estruturais determinam a biosíntese de peptídeos em proteínas funcionais, bem como os mecanismos das interações proteína-proteína e proteína-ligante utilizando algoritmos complexos. Empresas farmacêuticas e de biotecnologia utilizam a bioinformática na descoberta de medicamentos, para a obtenção de novas drogas específicas para o tratamento das doenças e que causem mínimos danos sistêmicos aos pacientes (Fenstermacher, 2005). Em termos práticos, a bioinformática pode ajudar a responder perguntas como, por exemplo, se um gene recentemente descoberto e analisado é semelhante a algum outro gene previamente conhecido, se a sequência de determinada proteína pode sugerir sua função, ou ainda, se genes relacionados a uma célula cancerígena são diferentes daqueles encontrados em uma célula saudável (Franco et al., 2008). É impossível categorizar como a bioinformática influencia o vasto campo da biologia, mas estes exemplos demonstram a natureza diversa do que a bioinformática é atualmente e poderá vir a se tornar no futuro.

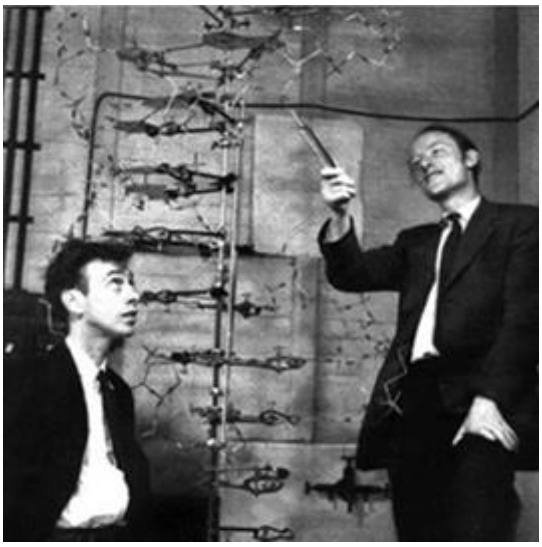
1.1 Introdução à Bioinformática

Embora os pioneiros da biologia computacional não tenham utilizado o termo "*bioinformática*" para descrever seus trabalhos, eles tinham uma clara visão de como a tecnologia da computação, matemática e biologia molecular poderiam ser proveitosamente combinadas para responder perguntas fundamentais das ciências da vida. A bioinformática é o resultado da união indissolúvel entre a tecnologia da informação e as ciências da vida, sendo originalmente destinada a resolver questões como: Como armazenar e organizar sequências de DNA? Como encontrar íntrons e exons em sequências de DNA genômico? Quais as condições necessárias para a transcrição de um gene em particular? Como aprender mais sobre a estrutura de uma proteína? Como comparar sequências proteicas previstas ou suas estruturas? Na era pós-genômica, a aquisição de novas e melhoradas ferramentas computacionais permitiu a bioinformática se tornar pivô de aplicações como o rastreamento genético, o diagnóstico molecular, a descoberta de drogas e o melhoramento genético de culturas (Franco et al., 2008).

Três fatores importantes facilitaram o surgimento da biologia computacional durante o início dos anos 60. Em primeiro lugar, a expansão da coleção de sequencias de aminoácidos forneciam uma fonte de dados e um conjunto interessante de problemas para resolver, o que seria impossível sem o poder de processamento de computadores. Em segundo lugar, a ideia de que macromoléculas carregam informações, o que se tornou parte essencial da estrutura conceitual da biologia molecular, provavelmente forneceu uma importante ligação entre a ciência da computação, a teoria da informação e a biologia molecular. Em terceiro lugar, os computadores digitais de alta velocidade, que se foram construídos durante a Segunda Guerra Mundial para o desenvolvimento de softwares de guerra, finalmente se tornaram disponíveis para biólogos em suas pesquisas acadêmicas. Nem todos os biólogos tinham – ou queriam ter – acesso a estas máquinas, mas, a partir de 1960, a escassez de computadores não era mais um obstáculo para o desenvolvimento da biologia computacional (Hagen, 2000).

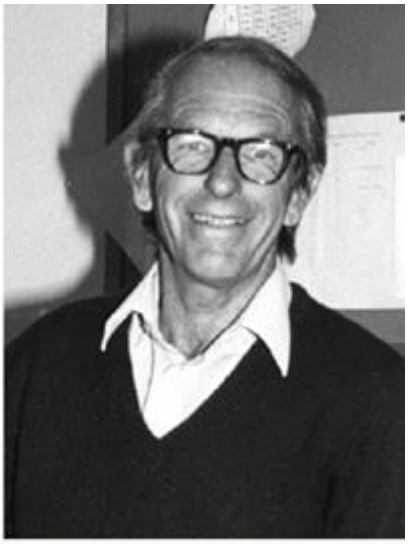
1. 1. 1 Contexto Histórico

Quando, em 1953, Watson e Crick propuseram o modelo de dupla hélice para explicar a estrutura do DNA, não imaginavam o volume exponencial de informações que seria gerado a partir deste momento. Com sorte, nas décadas seguintes, as ferramentas computacionais possibilitaram a análise e resolução de questões que foram criadas ao se desvendar a estrutura do DNA, como por exemplo, que a informação genética codifica para proteínas, as propriedades estruturais destas e seus fatores regulatórios, bem como eventos associados a regulação dos genes, bases moleculares do desenvolvimento embrionário e evolução de vias metabólicas e bioquímicas. Ao contrário do que se poderia esperar as ferramentas computacionais começaram a ser aplicadas na biologia molecular muito antes do início da Internet ou dos projetos de sequenciamento genômico (Franco et al., 2008).



Watson e Crick em frente a um modelo da hélice de DNA. *Cavendish Laboratory*, Universidade de Cambridge, 1953.

A ideia de que as proteínas podem transportar informações codificadas em sequencias lineares de aminoácidos é comum atualmente, porém esta é uma história relativamente recente. Esta teoria surgiu pela primeira vez durante as décadas seguintes à Segunda Guerra Mundial. Os estudos de Frederick Sanger (Ryle et al., 1955), que lhe renderam o Prêmio Nobel de Química em 1958, estabeleceram firmemente a teoria da estrutura polipeptídica das proteínas. Formulado primeiramente em 1902, este conceito havia enfrentado considerável ceticismo e concorrência com teorias alternativas. As técnicas de análise bioquímica de proteínas melhoraram muito durante os anos de 1930 e 1940, mas antes dos trabalhos de Sanger, não se sabia praticamente nada sobre a ordem de aminoácidos em qualquer proteína. Naquela época, estudiosos ainda se apegavam a crença de que proteínas eram estruturalmente simples ou até mesmo que não tinham uma estrutura definida. O divisor de águas deste período foi o sequenciamento completo da primeira proteína, a insulina, por Sanger e seus colaboradores, na



Frederick Sanger na cerimônia do Prêmio Nobel em 1980 (Hagen, 2000).

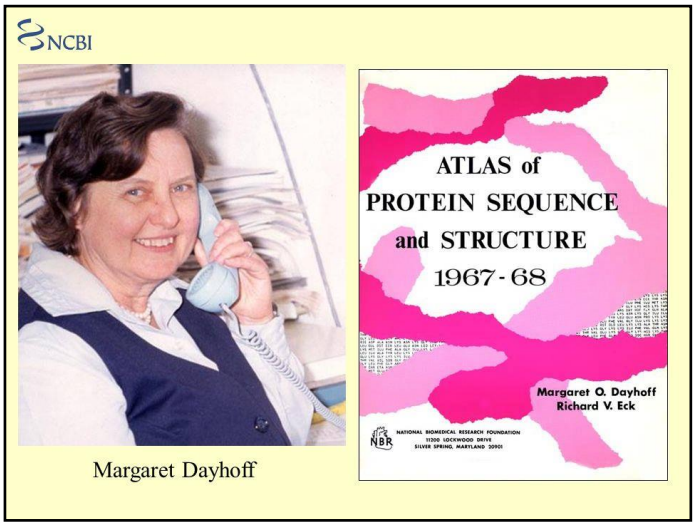
Universidade de Cambridge dos anos de 1945 a 1955 (Hagen, 2000).

Ao mesmo tempo, no entanto, outros bioquímicos estavam desenvolvendo métodos mais refinados que transformariam o processo analítico trabalhoso utilizado por Sanger e seus colaboradores. A reação de degradação de Edman, através da qual bioquímicos podem remover e identificar aminoácidos individuais sequencialmente a partir da região amino terminal de um peptídeo curto, representou uma grande melhoria em relação aos métodos descritos por Sanger (Fruton, 1992). O uso de colunas de troca iônica e outras inovações na cromatografia e eletroforese também tornaram o sequenciamento mais eficiente. Assim, como consequência, rapidamente todo o processo de separação e identificação de ácidos nucléicos foi tornando-se automatizado. Além disso, as técnicas semi-automatizadas utilizadas por pesquisadores liderados por Stanford Moore e William Stein no Instituto Rockefeller, eram capazes de sequenciar 124 aminoácidos de uma ribonuclease em metade do tempo em que o grupo de Sanger precisou para desvendar a sequência de 51 aminoácidos da insulina. A automatização provocou um choque na comunidade bioquímica, pois prometia transformar a realização do sequenciamento em um procedimento de rotina, não necessitando ser executado por grandes mestres da química, mas por qualquer técnico de laboratório competente. No final dos anos 1960, Pehr Edman projetou o primeiro "sequenciador", uma máquina de sequenciamento automatizada baseada em sua reação de degradação já amplamente utilizada na época (Edman & Begg, 1967). Tais inovações encorajaram muitos laboratórios a começarem trabalhos envolvendo o sequenciamento de proteínas, o que rapidamente aumentou a biblioteca de sequencias de aminoácidos (Hagen, 2000).

Nas mesmas décadas, foram publicados os primeiros estudos que elucidaram muitas questões sobre a estrutura das proteínas. Os trabalhos de Robert Corey, no início da década de 1950, e de Gopalasamudram N. Ramachandran, nos idos de 1960, que ofereceram as bases para a compreensão da estrutura tridimensional de proteínas (Verli, 2014). Contudo, os dados de sequência desempenharam um papel fundamental na interpretação das imagens de difração de raios-X utilizados por John Kendrew e Max Perutz quando determinaram as

Primeiro programa de visualização da estrutura 3D de moléculas, em fotografia publicada na revista *Scientific American*, em 1966.

estruturas tridimensionais de mioglobina (Kendrew, et al., 1958; Kendrew et al., 1960) e hemoglobina (Perutz, 1960). Combinar as técnicas bioquímicas de análise de sequência com as técnicas biofísicas de cristalografia de raios-X parecia ser a chave para a compreensão de como a informação molecular em uma sequência de aminoácidos promove o elevado grau de complexidade do dobramento de proteínas em uma configuração tridimensional específica (Hagen, 2000). Desde estes trabalhos até a primeira vez em que se relatou o uso de programas de computador para a visualização de estruturas tridimensionais de moléculas, passaram-se mais alguns anos quando, em 1966, foi publicado por Cyrus Levinthal, na revista *Scientific American*, o trabalho desenvolvido no *Massachussetts Institute of Tecnology* por John Ward e Robert Stotz, demonstrando o uso de um programa de computador para a visualização de estruturas tridimensionais de proteínas. Ainda nesta década, no ano de 1965, o "*Atlas of Protein Sequence and Structure*", organizado por diversos autores, entre os quais se destaca Margaret Dayhoff, consistiu no primeiro esforço para a sistematização do conhecimento da estrutura tridimensional dos efetores da informação genética, as proteínas (Verli, 2014).



Margaret Dayhoff

Margaret Dayhoff, é considerada a "fundadora da bioinformática", exerceu um papel fundamental sobre o que entendemos hoje sobre a bioinformática, tanto por suas contribuições em relação ao alinhamento de sequencias quanto ao estudo da estrutura de proteínas. Foi uma das pioneiras no uso de computadores para o estudo de biomoléculas, incluindo tanto ácidos nucleicos quanto proteínas. Ela propôs o código de uma letra para a representação cada aminoácido ao invés das usuais três letras, em uma época da computação em que os dados eram armazenados em cartões perfurados, revolucionou a análise de dados biológicos, o que é amplamente utilizado até hoje. Desenvolveu as primeiras matrizes de substituição e fez importantes contribuições no desenvolvimento dos estudos filogenéticos. Também teve participação importante no desenvolvimento de métodos para o estudo de moléculas por cristalografia de raios-X (Verli, 2014).

Em 1977, o primeiro genoma de um organismo foi sequenciado, o vírus Φ-X174 (Sanger et al., 1977), outras milhares de sequências de DNA já haviam sido decodificadas e armazenadas em bases de dados. Com a quantidade crescente de informações, a análise de sequencias de DNA

manual já se tornara impraticável. Emerge então um consenso de que era necessário um banco internacional de ácidos nucleicos e em 1979, em um workshop realizado pela *National Science Foundation* na Universidade Rockefeller é emitido um chamado para a criação dessa base de dados, nos dois anos seguintes foram realizadas uma série de oficinas para definir o projeto que culminou em 1982, com o início oficial do GenBank (Cravedi, 2008). No ano de 1990 o *National Institutes of Health* (NIH) e o *Department of Energy* (DOE) se juntam à parceiros por todo o mundo para iniciar o Projeto Genoma Humano, (HGP, do inglês *Human Genome Project*). Em 1995 ocorre o mapeamento da primeira bactéria, a *Haemophilus influenzae* Rd, todas as suas 1.830.137 pares de bases de nucleotideos foram apresentadas no trabalho de Fleischmann e colaboradores (1995).

O HGP foi oficialmente iniciado nos Estados Unidos em 1990, teve o envolvimento de mais de 5000 cientistas, de 250 diferentes laboratórios em todo o mundo, teve um investimento de mais de 3 bilhões de dólares e demorou 15 anos para ser concluído. O HGP foi um consórcio internacional, 17 países iniciaram programas de pesquisa sobre o genoma humano. Os maiores programas desenvolvem-se na Alemanha, Austrália, Brasil, Canadá, República Popular da China, Coreia do Sul, Dinamarca, Estados Unidos, França, Israel, Itália, Japão, México, Reino Unido, Rússia e Suécia, e ainda, outros países não citados aqui, mas que contribuíram com estudos envolvendo técnicas de biologia molecular de aplicação da pesquisa genética e de organismos de interesse particular para suas regiões geográficas. Este consórcio publicou um esboço inicial na revista científica *Nature* em fevereiro de 2001 com cobertura de cerca de 90 por cento do genoma (Venter et al., 2001).

Mais de uma década após a conclusão do HGP pudemos observar o crescimento sem precedentes no volume de dados biológicos. Devido ao avanço biotecnológico em tecnologias de alto rendimento, como *Microarray* e *Next Generation Sequencing*, atualmente é possível produzir dados biológicos de alta qualidade e a uma rápida velocidade. Descobertas, feitas a partir de dados biológicos, podem levar a uma melhor compreensão dos mecanismos das doenças e mais, orientar para um melhor diagnóstico e terapia (Raza, 2015). Neste contexto, a bioinformática é peça crucial. Não seria possível analisar um volume tão expressivo e crescente de dados se não fossem as ferramentas computacionais atualmente disponíveis, além disso, técnicas de bioinformática são atualmente empregadas para a descoberta de novos medicamentos, a uma demanda muito menor de custos e tempo de investigação

2 Bancos de Dados

2.1 Introdução

Quando Sanger descobriu o primeiro método para sequenciar proteínas, o interesse inicial em bioinformática foi impulsionado pela necessidade de criar bases de dados de sequências biológicas. O primeiro banco de dados foi criado anos após a disponibilização da primeira sequência de uma proteína em 1956, a insulina, que consistia em apenas 51 resíduos de aminoácidos (análogos aos caracteres alfabéticos contidos em apenas uma frase) (Babu, 1997). Em 1965, o *Atlas of Protein Sequence and Structure*, uma publicação anual que tentou catalogar todas as sequencias de aminoácidos conhecidas, e, embora rudimentar para os padrões de hoje, serviu como o primeiro banco de dados para a biologia molecular, tornando-se um recurso indispensável para a investigação computacional desde então. Mais tarde, evoluiu para um grande banco de dados *on-line*, o *Protein Information Resource (PIR)*, criado em 1983, e forneceu um importante ponto de partida para que outros biólogos computacionais logo começassem a construir suas próprias bases de dados moleculares (Hagen, 2000). Em meados dos anos 70 a primeira sequência de um ácido nucleico, o tRNA de uma levedura, com 77 bases (unidades individuais de ácidos nucleicos) foi descoberta. Durante este período, a estrutura tridimensional de proteínas foi estudada e o *Protein Data Bank (PDB)* foi desenvolvido, como o primeiro banco de dados de estruturas de proteínas, com apenas 10 entradas, em 1972. Hoje em dia, o PDB é uma grande base de dados com mais de 120.000 estruturas depositadas (PDB, 2016). Enquanto que as primeiras bases de dados de proteínas eram mantidas em laboratórios individuais, o desenvolvimento de uma base de dados oficial foi consolidado com o *Swiss-Prot*, um banco de dados de sequencias de proteínas iniciado em 1986 (Babu, 1997).

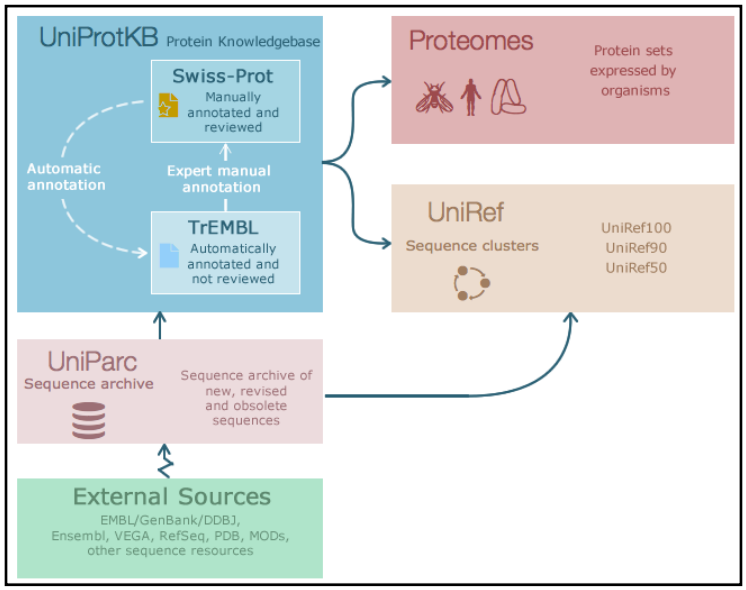
Nas últimas décadas pudemos observar a aplicação intensa da tecnologia computacional para a análise e modelagem de dados biológicos. Atualmente, isso exerce um enorme impacto sobre o entendimento e desenvolvimento da biologia molecular. Um aspecto importante desta revolução é o armazenamento, recuperação e análise de conjuntos de dados biológicos (Mathura & Kangueane, 2009). Além disso, a disponibilidade de grandes quantidades de dados biológicos públicos e privados demanda a necessidade constante de transformar estas informações em conhecimentos úteis. Entender as correlações, estruturas e padrões dos dados biológicos são as mais importantes tarefas da bioinformática. A informação e o entendimento nas múltiplas áreas de conhecimento podem ser então utilizados para aplicações que incluem, por exemplo, a descoberta de medicamentos, análise de genomas e controle biológico (Chen, 2005).

A partir da década de 1990, com os processos automatizados de sequenciamento em larga escala, foi necessária a construção de bancos de dados mais robustos para abrigar a explosão no número de sequências obtidas pelos pesquisadores. O NCBI, por exemplo, foi criado pelo NIH (*National Institutes of Health*, o Instituto Nacional de Saúde dos Estados Unidos) em 1988 para abrigar esse tipo de informação (Wheller et al., 2002). Dessa forma, foi criada uma colaboração internacional para montar um banco de dados de sequências de nucleotídeos, a INSDC (*International Nucleotide Sequence Database Colaboration*). Essa instituição contém o NCBI, o EMBL (*European Molecular Biology Laboratory*) e o DDBJ (*DNA Data Bank of Japan*) (Tateno et al., 2002). Cada um desses centros possibilita a submissão individual de sequências de DNA e trocam informações entre si diariamente, sendo que todos os três possuem informações atualizadas de todas as sequências disponíveis para os pesquisadores (Stoesser et al., 2002). Apesar disso, cada centro apresenta os dados de forma particular, apesar de bastante semelhante. Ultimamente têm surgido uma grande quantidade de novos bancos de dados em biologia molecular. E são tantos, que uma das principais revistas da área, a inglesa *Nucleic Acids Research* (http://nar.oupjournals.org/), tem reservado dois números especiais por ano (os primeiros volumes dos meses de janeiro e julho) para apresentar apenas artigos sobre novos bancos de dados ou de atualizações de bancos já consagrados pela comunidade (Prosdocimi, 2007).

Os bancos de dados biológicos consistem em um conjunto de informações inter-relacionadas, organizadas, acessíveis, gerenciáveis e atualizadas constantemente. Além disso, as ferramentas atualmente disponíveis permitem ao usuário além da pesquisa e consulta de informações, cruzar e correlacionar de dados seu interesse. Bem como a disponibilidade de técnicas de alto rendimento, os bancos de dados biológicos vêm sendo gerados de forma exponencial e a biologia moderna transformou-se em uma ciência rica em dados. Alguns dados biológicos importantes e úteis disponíveis nos bancos de dados são sequências de nucleotídeos e de proteínas, estruturas tridimensionais de proteínas resolvidas através da cristalografia de raios-X e ressonância magnética nuclear (RMN), vias metabólicas, mapas e genomas completos, expressão de genes e interação proteína-proteína, etc. Os bancos de dados biológicos podem ser amplamente divididos em bancos de dados de sequencias e de estruturas. Os dados de sequência podem ser aplicáveis tanto a estudos genômicos quanto de proteínas, mas as bases de dados estruturais são aplicáveis apenas para proteínas. Hoje, a maioria destes, estão disponíveis gratuitamente para os pesquisadores. Em geral, os bancos de dados biológicos podem ser classificados como primários e secundários (Raza, 2015).

2. 2 Bancos de Dados Primários

Os bancos de dados primários são obtidos a partir de informações experimentais como o sequenciamento, difração de raios-X ou ressonância magnética nuclear (RMN). Pode conter anotações como sítios de fosforilação de proteínas, regiões promotoras de genes, entre outras. Os dados são submetidos diretamente pelos pesquisadores, os próprios autores são responsáveis e controlam os conteúdos que submetem. Ex.: GenBank, EBI (EMBL), DDBJ, Uniprot, PDB.



O consórcio **UniProt** (*Universal Protein Resource*) compreende os institutos *European Bioinformatics Institute* (**EBI**), *Swiss Institute of Bioinformatics* (**SIB**) e *Protein Information Resource* (**PIR**). O EBI, localizado no Campus *Wellcome Trust Genome* em Hinxton, Reino Unido, abriga muitos recursos de bases de dados e serviços de bioinformática. O SIB, localizado em Genebra, Suíça, mantém os servidores ExPASy (*Expert Protein Analysis System*), que são um recurso central para se obter ferramentas proteômicas e bancos de dados. O **PIR**, organizado pela *National Biomedical Research Foundation* (NBRF) no Centro Médico da Universidade de Georgetown em Washington, é o herdeiro do mais antigo banco de dados de sequências de proteínas, o *Atlas of Protein Sequence and Structure* de Margaret Dayhoff, publicado pela primeira vez em 1965. O **EMBL-EBI** e o **SIB** foram utilizados em conjunto para produzir o **Swiss-Prot** e **TrEMBL**, enquanto o PIR produziu o *Protein Sequence Database* (PIR-PSD). Estes dois conjuntos de bancos de dados conviveram com diferentes prioridades de cobertura

sequências de proteínas e anotação. O **TrEMBL** (Traduzido EMBL - Banco de dados de sequência de nucleotídeos) foi originalmente criado porque dados de sequenciamento estavam sendo gerados em um ritmo que ultrapassou a capacidade de manutenção do Swiss-Prot. Enquanto isso, o PIR manteve o **PIR-PSD** e bancos de dados relacionados, incluindo **iProClass**, um banco de dados de curadoria de sequências e famílias de proteínas. Em 2002, a EBI, SIB, e PIR juntaram forças como o consórcio UniProt.

As bases de dados do UniProt são *Knowledgebase UniProt* (**UniProtKB**), o *UniProt Reference Clusters* (**UniRef**) e o UniProt Achive (**UniParc**).

GenBank

O GenBank é um banco de dados de sequência de acesso livre, todas as anotações da coleção de nucleotídeos, sequências e traduções de proteínas são publicamente disponíveis. Esta base de dados é mantida e produzida pelo *National Center for Biotechnology Information* (NCBI), como parte do *International Nucleotide Sequence Database Collaboration* (INSDC). O NCBI é parte do *National Institutes of Health* – NIH (Instituto Nacional de Saúde) dos Estados Unidos. O GenBank e seus colaboradores recebem sequencias produzidas em laboratórios de todo o mundo, de mais de 100.000 organismos diferentes. Nos mais de 30 anos desde sua criação, o GenBank tornou-se a base de dados mais importante e influente para a pesquisa em quase todos os campos biológicos, cujos dados são acessados e citados por milhões de pesquisadores no mundo. Este banco de dados continua a crescer a uma taxa exponencial, dobrando suas entradas a cada 18 meses. O GenBank é produzido através do envio direto de sequencias por parte de laboratórios de pesquisa individuais, bem como do recebimento de dados em massa a partir de centros de sequenciamento em larga escala.

2. 3 Bancos de Dados Secundários

Os bancos de dados secundários são compostos por informações derivadas dos bancos de dados primários, porém seu conteúdo é controlado por curadores (EBI, SIB). Ex.: Blocks, Smart, Print, PFam, Prosite, Conserved Domain, etc.

2. 4 Exercícios:

1) Acesse: http://www.ncbi.nlm.nih.gov/genbank/

✓ Buscar no GenBank a sequência do RNAm que codifica para a proteína ***Kin17 humana*** no formato ***FASTA***:

```
>gi|3850703|emb|AJ005273.1| Homo sapiens mRNA for Kin17 protein
CTAGAATTCAGCGGCCGCTGAATTCTAGAACTGGGGTCCAGAAAGTGATCGCTGCCGTGGTCGCCATGGG
GAAGTCGGATTTTCTTACTCCCAAGGCTATCGCCAACAGGATCAAGTCCAAGGGGCTGCAGAAGCTACGC
TGGTATTGCCAGATGTGCCAGAAGCAGTGCCGGGACGAGAATGGCTTTAAGTGTCATTGTATGTCCGAAT
CTCATCAGAGACAACTATTGCTGGCTTCAGAAAATCCTCAGCAGTTTATGGATTATTTTTCAGAGGAATT
CCGAAATGACTTTCTAGAACTTCTCAGGAGACGCTTTGGCACTAAAAGGGTCCACAACAACATTGTCTAC
AACGAATACATCAGCCACCGAGAGCACATCCACATGAATGCCACTCAGTGGGAAACTCTGACTGATTTTA
CTAAGTGGCTGGGCAGAGAAGGCTTGTGCAAAGTGGACGAGACACCAAAAGGCTGGTATATTCAGTACAT
AGACAGGGACCCAGAAACTATCCGCCGGCAACTGGAACTGGAGAAAAAGAAAAAGCAGGACCTTGATGAT
GAAGAAAAAACTGCCAAATTTATTGAAGAGCAAGTGAGAAGAGGCCTGGAAGGGAAGGAACAGGAGGTCC
CTACTTTTACGGAATTAAGCAGAGAAAATGATGAAGAGAAAGTCACGTTTAATTTGAGTAAAGGAGCATG
TAGCTCATCCGGAGCAACATCTTCCAAGTCAAGTACTCTGGGACCGAGTGCACTGAAGACGATAGGAAGT
TCAGCATCAGTGAAACGAAAAGAATCTTCCCAGAGCTCAACTCAGTCTAAAGAAAAGAAGAAAAAGAAAT
CTGCACTGGATGAAATCATGGAGATTGAAGAGGAAAAGAAAAGAACTGCCCGAACAGACTACTGGCTACA
GCCTGAAATTATTGTGAAAATTATAACCAAGAAACTGGGAGAGAAATATCATAAGAAAAAGGCTATTGTT
AAGGAAGTAATTGACAAATATACAGCTGTTGTGAAGATGATTGATTCTGGAGACAAGCTGAAACTTGACC
AGACTCATTTAGAGACAGTAATTCCAGCACCAGGAAAAAGAATTCTAGTTTTAAATGGAGGCTACAGAGG
AAATGAAGGTACCCTAGAATCCATCAATGAGAAGACTTTTTCAGCTACTATCGTCATTGAAACTGGCCCT
TTAAAAGGACGCAGAGTTGAAGGAATTCAATATGAAGACATTTCTAAACTTGCCTGAGTTTGAAAATTTG
TTAACAATACATTAAAATCTTAAAGCATCAAATTGGTGTTCGCCAAGGCATTATGAGACTCTACTGTGTT
AGGGTATATTCTTTTGTATAAAACAAACAGGTTTTTGAAAATATTACTGTATAGTTGTTCAGCTAAACTT
TGAGAAGAATTTAATTATGTCTCATGAGGTATCAAACTATGTAATTTTGTCCTTGTTATTTTTGTTTCCT
TTGTAATTTACTTGATGAGTTTATATCTTCATTAAAGAATGTTATTATAAAAAAAAAA
```

* Em bioinformática, ***FASTA*** é um formato baseado em texto para representar tanto sequências de nucleótidos quanto de aminoácidos, no qual os nucleotídeos ou aminoácidos são representados usando códigos de uma única letra. A linha de descrição se distingue a partir da sequência dos dados por um símbolo maior-que (">") na primeira coluna. A palavra que segue o símbolo ">" é o identificador da sequência, e o resto da linha é a descrição (ambos são opcionais). Não deve haver nenhum espaço entre o ">" e a primeira letra do identificador.

2) Acesse: http://www.uniprot.org/.

a) Utilize a mesma sequência do RNAm utilizada na busca anterior e faça uma análise da região codificante através do servidor Expasy (http://web.expasy.org/translate/).

b) Cole a sequência do RNAm no espaço em branco;

c) Clique no botão e TRANSLATE SEQUENCE aguarde o resultado;

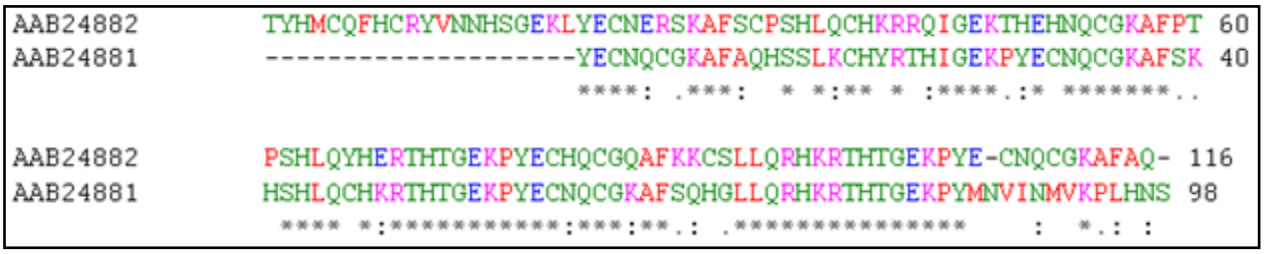
d) Analisa as ORFs e clique sobre a que coce acredita ser a que corresponde a região codificante da Kin17;

e) Clique na primeira Metionina e veja quantos aminoácidos a proteína possui;

f) Vá até o Uniprot e confirme sua análise procurando pela sequência da Kin17 humana.

2.5 Ferramentas Para Alinhamento de Sequências

Alinhamento de sequências é uma forma de organizar sequências primárias de DNA, RNA ou proteínas para identificar regiões similares que possam ser consequência de relações funcionais, estruturais ou evolucionárias entre elas. Sequências alinhadas de nucleotídeos ou resíduos de aminoácidos são representadas tipicamente como linhas de uma matriz. Espaçamentos (*gaps*) podem ser inseridos entre os resíduos para que caracteres semelhantes (por algum critério) sejam alinhados em colunas sucessivas.



Exemplo de alinhamento entre duas sequências, produzido pelo programa ClustalW entre duas proteínas dedo-de-zinco humanas (human zinc finger proteins) identificadas por seus números de acesso no GenBank.

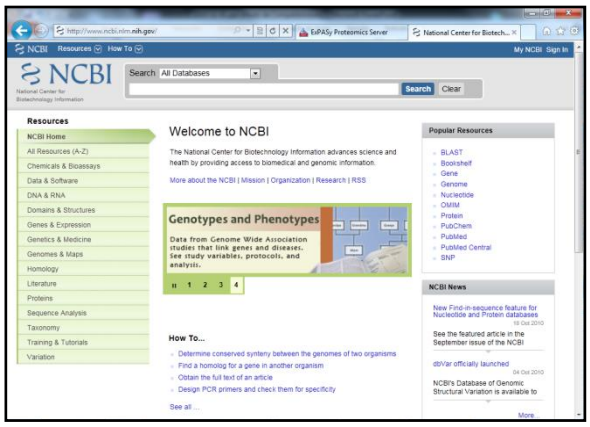
Se duas sequências em um alinhamento compartilham de um ancestral comum, discordâncias (*mismatches*) podem ser interpretadas como mutações pontuais e os espaços (*gaps*) como inserções ou deleções introduzidas em uma ou ambas as sequencias desde quando estas divergiram no tempo.

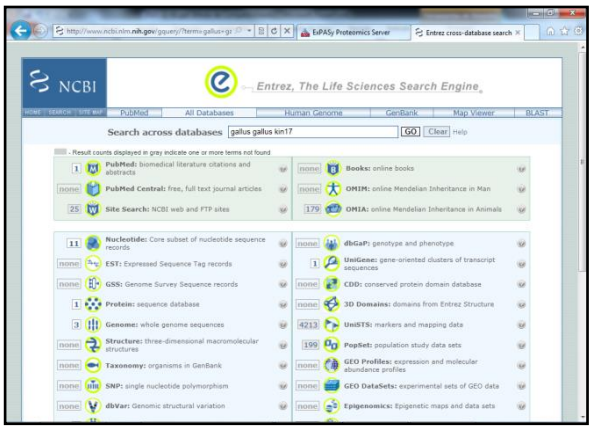
Abordagens computacionais para o alinhamento de sequências dividem-se, em geral, em duas categorias: alinhamentos globais e alinhamentos locais. Calcular um alinhamento global é uma forma de otimização global que "força" o alinhamento a cobrir todo o comprimento de todas as sequencias interrogadas (*query*). Por outro lado, os alinhamentos locais identificam regiões de similaridade dentro de sequencias longas que são geralmente bastante divergentes em um todo. Os alinhamentos locais são frequentemente preferíveis, mas podem ser difíceis de calcular por causa do problema adicional de identificar regiões internas de similaridade. Uma grande variedade de algoritmos existem para abordar o problema de alinhamento de sequencias, sendo os mais conhecidos os baseados em programação dinâmica, mais lentos porém teoricamente otimizadores, ou baseados em heurística, mais eficientes/rápidos mas sem prova formal de obtenção de solução ótima.

a) Busque nas bases de dados Uniprot as sequências da proteína humana Kin17 ou outras de seu interesse.
b) Copie e cole as sequências no formato FASTA usando o WordPad.
c) Alinhe as sequências usando o UniprotKB/**Alignment**
d) Explore as similaridades entre elas selecionando as propriedades dos aminoácidos.

✓ Acesse (http://www.ncbi.nlm.nih.gov/)**)**

O *National Center for Biotechnology Information* (**NCBI**) foi fundado em 1988 e é parte da *United States National Library of Medicine* **(NLM**), uma filial do *National Institutes of Health* (**NIH**) e está localizado em Bethesda, Maryland, USA.





2.6 BLAST (Basic Local Alignment Search Tool)

BLAST (sigla em inglês que significa: *Basic Local Alignment Search Tool*), é um algoritmo para comparar informações de sequências biológicas primárias, tais como seqüências de aminoácidos de diferentes proteínas ou nucleotídeos de seqüencias de DNA.

Uma pesquisa BLAST permite que um investigador compare uma sequência fornecida em uma consulta com uma biblioteca ou base de dados de sequências, identificar as bibliotecas de seqüências que se assemelham a aquela consultada e que estejam acima de um certo grau de semelhança.

Numa situação hipotética, após descobrir um gene anteriormente desconhecido em um camundongo, um cientista poderia tipicamente elaborar uma pesquisa no BLAST do genoma humano para verificar se existem seres humanos portadores de um gene semelhante.

Diferentes tipos de BLAST

BLASTN – Pesquisa bancos de dados de nucleotídeos usando uma sequência de nucleotídeos em questão.
Algorítmos: blastn, megablast, discontiguous megablast

BLASTP – Pesquisa bancos de dados de proteínas usando uma Sequência de proteína em questão.
Algorithms: blastp, psi-blast, phi-blast

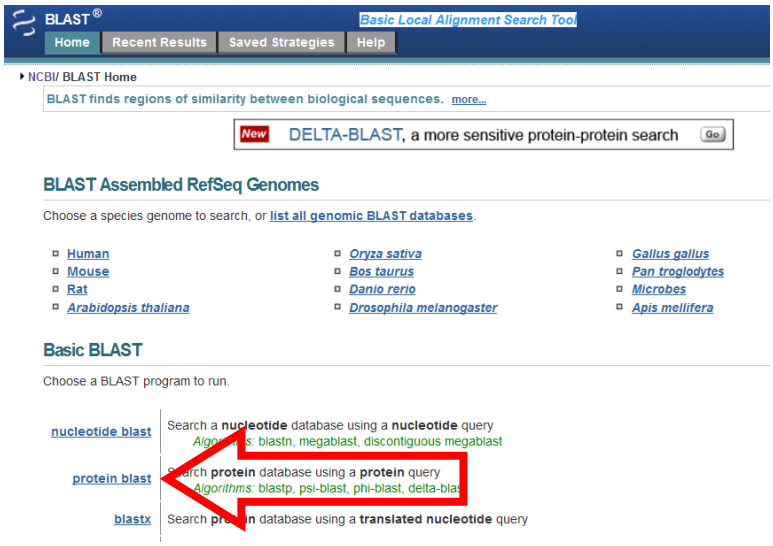
BLASTX – Pesquisa banco de dados de proteínas usando uma sequência de nucleotídeo traduzida em questão

TBLASTN – Pesquisa um banco de dados de nucleotídeos usando uma proteína em questão.

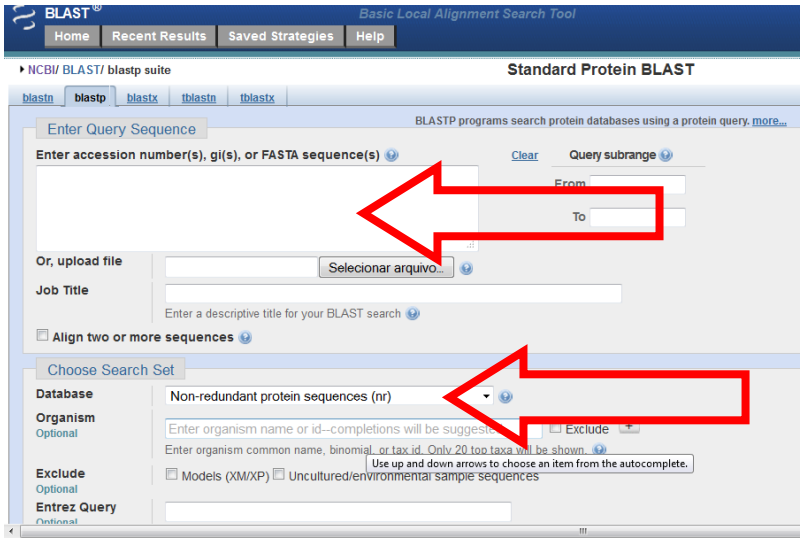
TBLASTX – Pesquisa bancos de dados de nucleotídeos traduzidos Usando um nucleotídeo traduzido em questão.

a) A partir da sequência de aminoácidos da Kin 17 humana, fazer uma busca por proteínas similares (moldes), cujas estruturas estejam depositadas no *Protein Data Bank*.

1º passo: Selecione a opção *Protein blast*

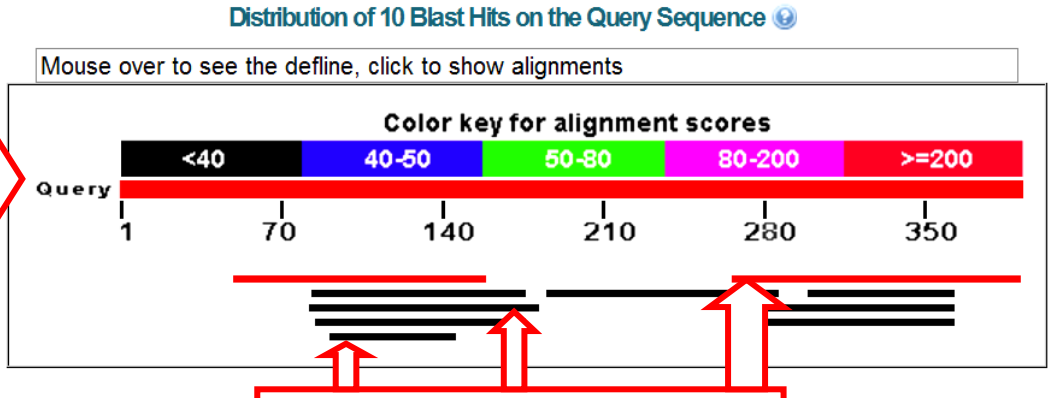


2º Passo: Cole a sequência no formato FASTA, selecione em seguida a opção *"Protein Data Bank proteins (PDB)"* em *"Database"*, clique em *"BLAST"* e aguarde o resultado.



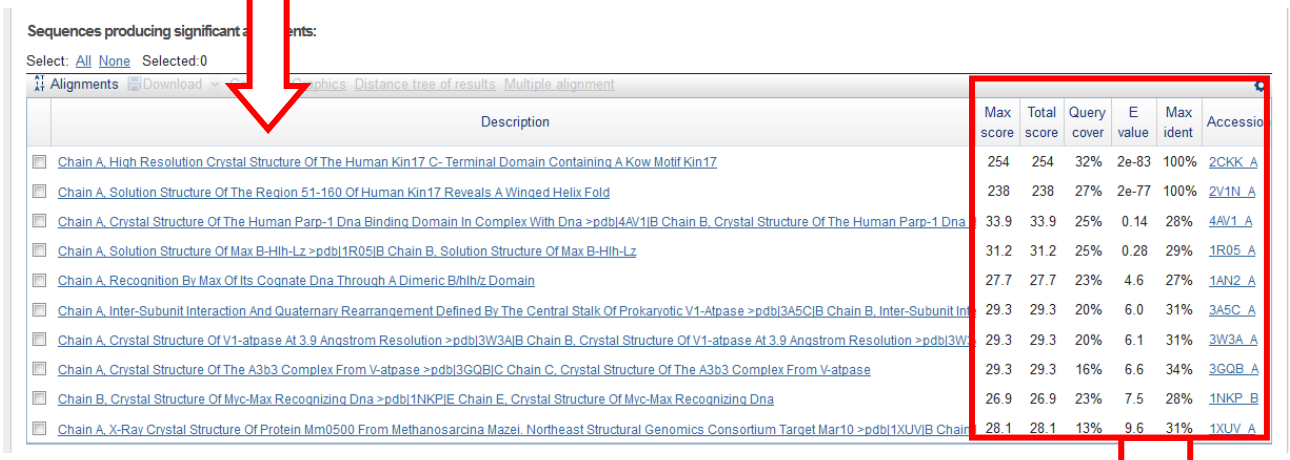
3º Passo: Analise o resultado.

Essa barra em vermelho representa a sequência que foi usada na pesquisa. Esta sequência é chamada de "*query*".



Cada uma dessas barras coloridas indicam a região onde a sequência do banco de dados bate com a sequência *query*

Descrição da proteína depositada no PDB por ordem de "*query coverage*"



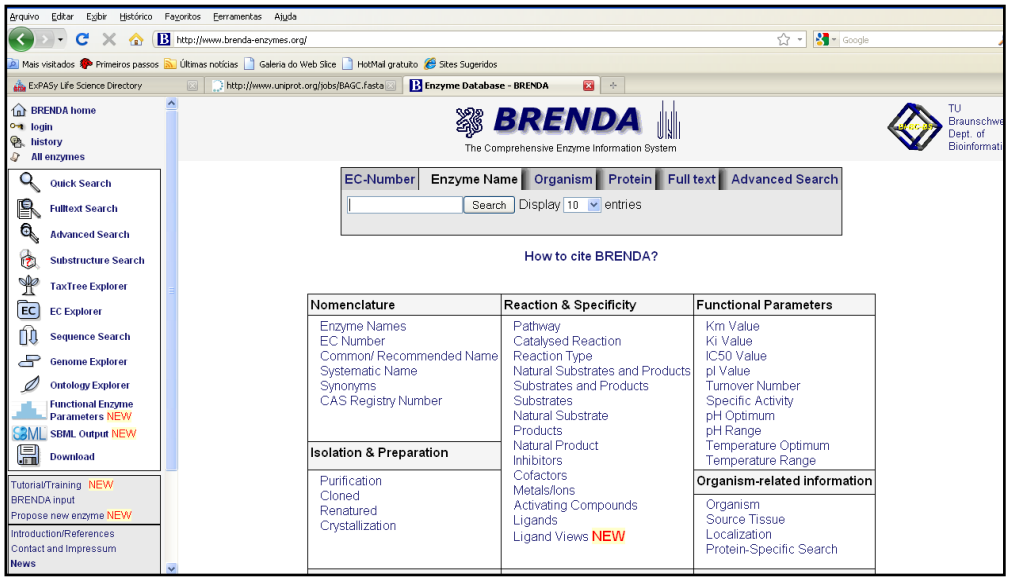
- ***Max score*:** Indica o quão bem sua sequência se encaixa;
- ***Total Score*:** inclui porções não contíguas da sequência do molde que corresponde a sequência *query*;
- ***Query coverage*:** Indica a fração da sequência *query* que corresponde a sequência do molde,
- ***E-value*:** Score que determina a confiabilidade do resultado, Quanto menor o escore, mais significativo é o alinhamento;
- ***Max ident.***: Refere-se a semelhança entre as sequências de aminoácidos das proteínas.

Objetivo: Prever o padrão de enovelamento (*folding*) da estrutura secundária a partir da sequência de aminoácidos.

✓ Usando as ferramentas expasy/**PSIPRED** insira a sequência da proteína de interesse no formato Fasta. Interprete os resultados.

BRENDA é o principal banco de dados de dados funcionais de enzimas, disponível para a comunidade científica. Possui um acervo com mais de 5.000 enzimas diferentes. É mantido e desenvolvido pelo **Instituto de Bioquímica e Bioinformática** da Universidade Técnica de Braunschweig, na Alemanha. É um repositório de dados sobre a função da enzima são extraídos diretamente da literatura primária por cientistas formados em Biologia ou Química. As verificações de forma e consistência são feitas por programas de computador. Cada conjunto de dados sobre uma enzima classificada é verificado manualmente por pelo menos um biólogo e um químico.

➢ Acesse (http://www.brenda-enzymes.org).



a) Buscar o ***E.C.***, o K_M, o ***pH ótimo*** e ***temperatura ideal*** da catalase bovina (*Bos taurus*) e de fungo (*Aspergillus. agaricus*), na sua forma livre.

b) Busque estas informações para outras enzimas de seu interesse.

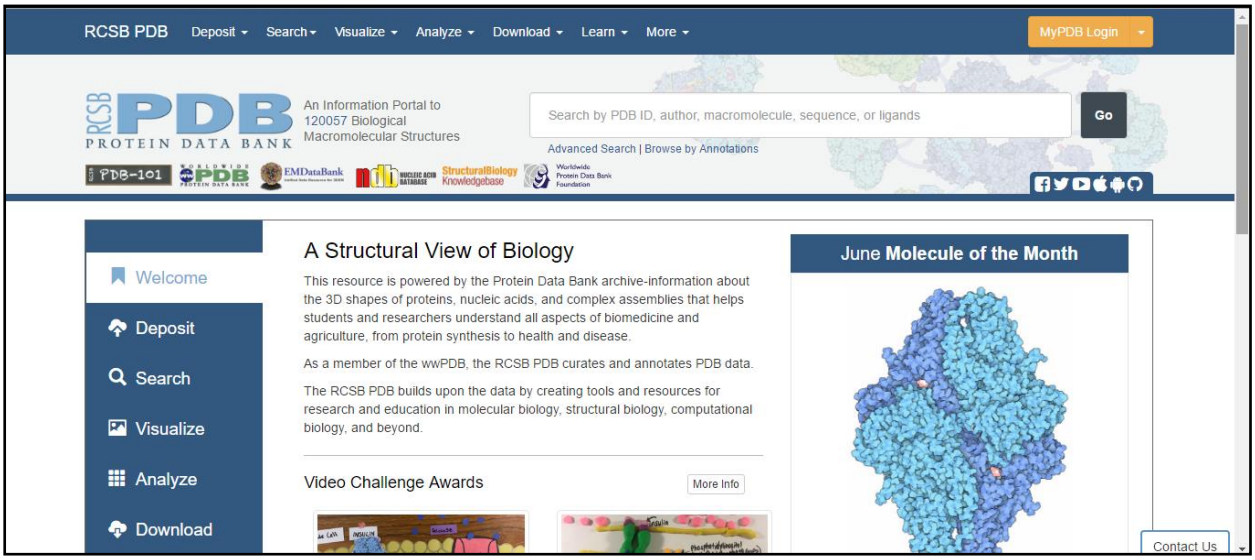
2.7 Outros repositórios de dados relacionados a proteômica:

2. 7. 1 Banco de dados de estruturas 3D

O ***Protein Data Bank*** (**PDB**) é o único repositório de informações em todo o mundo sobre as estruturas 3D de macromoléculas biológicas, incluindo as proteínas e ácidos nucléicos. A seguir, veremos um breve histórico desta importante base de dados:

1971 – Início do PDB no *Brookhaven National Laboratory* com 7 estruturas.
1998 – O *Research Collaboratory for Structural Bioinformatics* (RCSB) se tornou responsável pela manutenção do PDB.
2010 – o RCSB PDB é membro do wwPDB, um esforço colaborativo do PDBe (UK), PDBj (Japão) e do BMRB (USA) para assegurar que o arquivo PDB seja global e uniforme.

Este banco de dados, somente aceita o depósito de estruturas determinadas experimentalmente por cristalografia de raios-X, NMR ou Microscopia eletrônica.



- ✓ Acesse (http://www.rcsb.org/pdb/home/home.do)
- ✓ Procure a estrutura de sua proteína (ou relacionadas) utilizando a ferramenta ***Advanced Search, Macromolecule name*** (HMGB1) ou baixe uma estrutura pelo identificador (*pdbid*): 1KD2 e 1HVC. Explore a estrutura utilizando o visualizador *Jmol*.

3 Estrutura Tridimensional de Proteínas

Trechos retirados do livro: ***Princípios de Bioquímica de Lehninger****. 5. Ed. Porto Alegre: Artmed, p. 530-545, 2011*

3.1 Introdução

Em biotecnologia, o genoma é toda a informação hereditária de um organismo que está codificada em seu DNA (ou, em alguns vírus, no RNA). Isto inclui tanto os genes como as seqüências não codificadoras. Cada gene codifica uma determinada sequência de uma cadeia polipeptídica, chamada de proteína. Esss são por sua vez, os "tijolos" para a construção dos seres vivos.

Proteínas controlam praticamente todos os processos que ocorrem em uma célula, exibindo uma quase infinita diversidade de funções. Subunidades monoméricas relativamente simples fornecem a chave da estrutura de milhares de proteínas diferentes. As proteínas de cada organismo, da mais simples das bactérias aos seres humanos, são construídas a partir do mesmo conjunto onipresente de 20 aminoácidos. Como cada um desses aminoácidos tem uma cadeia lateral com propriedades químicas características, esse grupo de 20 moléculas precursoras pode ser considerado o alfabeto no qual a linguagem da estrutura proteica é lida. Para gerar uma determinada proteína, os aminoácidos se ligam de modo covalente em uma sequência linear característica. O mais marcante é que as células produzem proteínas com propriedades e atividades completamente diferentes ligando os mesmos 20 aminoácidos em combinações e sequências muito diferentes. A partir desses blocos de construção, diferentes organismos podem gerar produtos tão diversos como enzimas, hormônios, anticorpos, transportadores, fibras musculares, proteínas das lentes dos olhos, penas, teias de aranha, chifres de rinocerontes, proteínas do leite, antibióticos, venenos de cogumelos e uma miríade de outras substâncias com atividades biológicas distintas.

3. 2 Aminoácidos e Proteínas

Os aminoácidos são compostos orgânicos de função mista: possuem ao menos uma carboxila e um grupo amina. Os 20 aminoácidos das proteínas são todos α-aminoácidos, quer dizer, o grupo amina característico está ligado ao carbono alfa. O que distingue um aminoácido do outro é a natureza do grupo R. Eles tem formas, tamanhos e características distintas (figura 1).

Os aminoácidos podem unir-se através de uma ligação entre o grupo amina de um e o grupo carboxila de outro, chamada ligação peptídica. O produto resultante é um peptídeo; se muitos aminoácidos se unem o produto é um polipeptídeo. Uma proteina é um polipeptídeo.

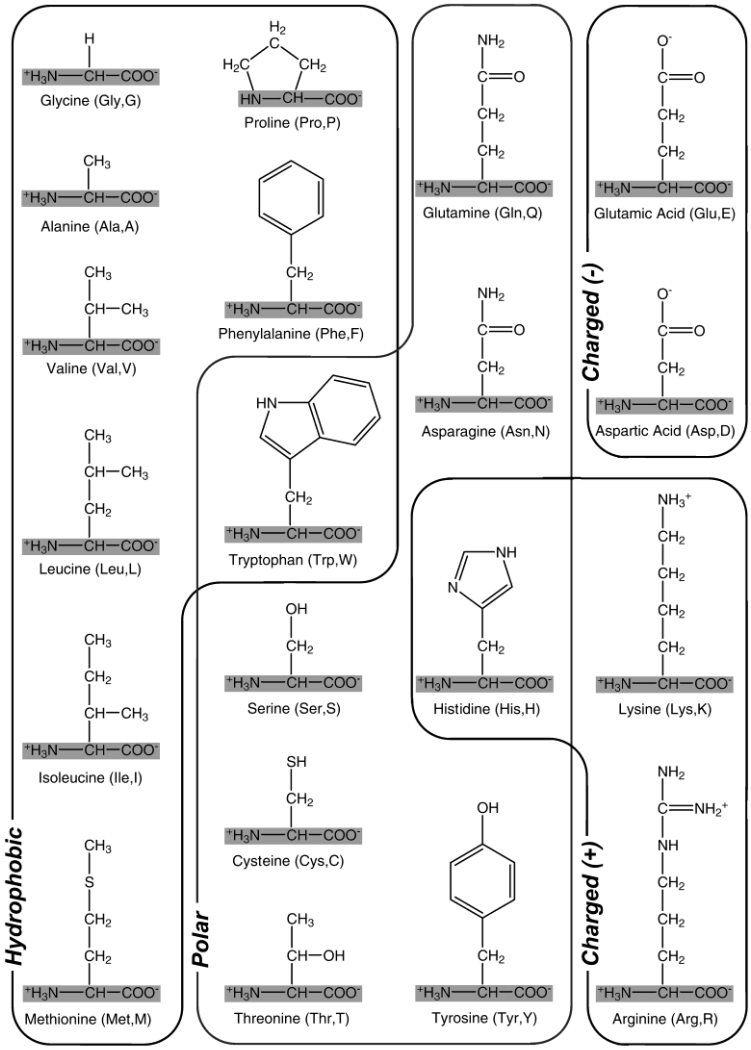


Figura 1. Os aminoácidos que compõem as proteínas divididos de acordo com seu grupo R.

A função de uma proteína está intimamente associada à sua estrutura tridimensional. Logo, as propriedades de cada aminoácido presente define o rumo de determinada proteína. Isso porque, existe uma série de interações fracas, em grande quantidade, que atuam na estrutura para que ela assuma a sua forma nativa. Para que a proteína assuma essa forma, basicamente existe um hierarquia estrutural que deve ocorrer, e recebem uma classificação: (Figura 2)

- ✓ **Estrutura primária:** é a sequência linear de aminoácidos que compõem uma proteína;
- ✓ **Estrutura secundária:** Padrão regular de pontes de hidrogênio nas proteínas que resultam em dois padrões que podem ser vistos em quase todas a estrutura de proteínas conhecidas: a α-*helice* e as *folhas-*β.

- ✓ **Estrutura terciária:** É definida como o arranjo tridimensional das estruturas secundárias numa cadeia polipeptídica. A estrutura terciária de uma proteína é determinada por interações não covalentes entre as cadeias laterais dos aminoácidos constituintes.
- ✓ **Estrutura quaternária:** Refere-se a organização das subunidades (cadeias polipeptídicas) em uma proteína de múltiplas subunidades. As subunidades podem ser idênticas ou diferentes.

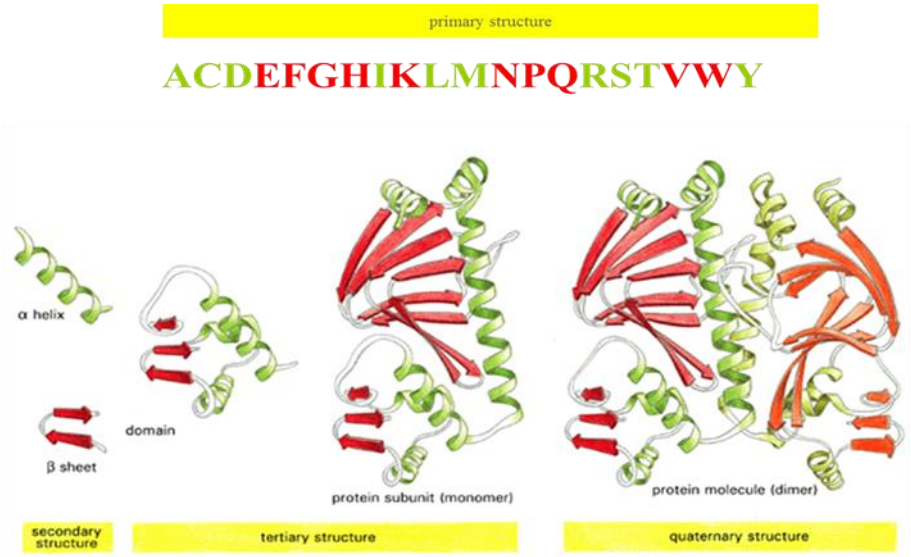


Figura 2. Classificação da estrutura das proteínas.

Quando a proteína assume sua forma nativa ela pode exercer sua função na célula. As proteínas estão presentes em todos os seres vivos e participam em praticamente todos os processos celulares, desempenhando um vasto conjunto de funções no organismo, como a replicação do DNA, a resposta a estímulos e o transporte de moléculas. Muitas proteínas são enzimas que catalisam reações bioquímicas vitais para o metabolismo. As proteínas têm também funções estruturais ou mecânicas, como é o caso da actina e da miosina nos músculos e das proteínas no citoesqueleto, as quais formam um sistema de andaimes que mantém a forma celular. Outras proteínas são importantes na sinalização celular, resposta imunitária e no ciclo celular.

Como vimos, as proteínas diferem entre si fundamentalmente na sua sequência de aminoácidos. Com isso, há uma necessidade em se conhecer como é cada estrutura entre os mais diversos organismos. Determinando-a obtemos dados, como localização de resíduos conservados e substituídos, sítios de ligantes, fendas/cavidades, relação evolucionária (filogenia) e Mecanismos de reação.

3. 3 Determinação experimental da Estrutura de proteínas

Visto que a estrutura é o ponto chave para a função de uma proteína em um organismo, a determinação desta é de extrema importância. Neste aspecto, várias técnicas são usadas, nas quais duas se destacam, que são a difração de raio-X e a ressonância magnética nuclear (RMN).

O processo para a obtenção da estrutura tridimensional de uma proteína via técnica de cristalografia por difração de raios-X é composto basicamente pela produção e purificação da proteína alvo, cristalização, coleta e processamento dos dados, resolução da estrutura (empregando informações sobre a sequência de aminoácidos e diferentes programas) e refinamento da estrutura. A técnica de RMN também requer o conhecimento da sequência de aminoácidos. Contudo, não é necessário que a proteína esteja em um estado de cristal ordenado. A vantagem da RMN é que a estrutura a ser determinada pode estar em solução, apesar de requerer que a proteína solubilizada esteja em altas concentrações. Infelizmente, esta técnica ainda está limitada a proteínas de tamanhos pequenos a médios, limitação não observada para a cristalografia. Mesmo assim, a RMN destaca-se ao revelar informações sobre o comportamento dinâmico das estruturas, incluindo mudanças conformacionais e interações com outras moléculas. Na RMN, um forte campo magnético alinha os momentos magnéticos dos núcleos atômicos de isótopos que possuem spin nuclear diferente de zero (tais como 1H, 13C, 15N, 9F e 31P). Uma fonte de radiofrequência de energia variável é emitida, podendo ser absorvida pelos núcleos atômicos invertendo o alinhamento do spin nuclear em relação ao campo magnético externo aplicado. Neste momento, parte da energia é absorvida e o espectro de absorção resultante fornece a informação sobre a identidade do núcleo e seu ambiente químico na vizinhança. Dados de sucessivos experimentos são coletados e um espectro de RMN é gerado contendo as informações sobre todos os deslocamentos químicos de todos os isótopos analisados na proteína.

A determinação experimental ainda é considerada o melhor processo para se obter a estrutura tridimensional de uma proteína. Entretanto estas técnicas, além de serem financeiramente custosas, podem levar anos e, em alguns casos, a estrutura final pode não chegar a ser obtida. Portanto, o desenvolvimento de métodos computacionais é tanto uma alternativa mais barata quanto, em alguns casos, a única possibilidade de obtenção de modelos estruturais para algumas proteínas.

4 Modelagem Molecular por Homologia

4.1 Introdução

A modelagem molecular por homologia é um método computacional que tem como princípio básico, modelar uma proteína utilizando uma proteína da mesma família. A proteína de interesse (alvo) terá sua estrutura 3D predita usando como referência a estrutura 3D de outra proteína similar (também chamada de molde, e na maioria das vezes evolutivamente relacionada). Essa proteína similar tem de possuir estrutura 3D resolvida experimentalmente, e as coordenadas cartesianas de seus átomos devem estar depositadas em banco de dados de estruturas como o Protein Data Bank (PDB). A modelagem molecular por homologia é o método empregado mais frequentemente, e seu limite de predição está intrinsecamente relacionado com o grau de similaridade entre as estruturas alvo e molde. Geralmente, consideram-se como limites mínimos de aplicabilidade do método valores de 25 a 30% de identidade, obtidos através do alinhamento entre a estrutura primária da proteína alvo e de uma ou mais proteínas molde. A modelagem por homologia pode ser dividida em cinco etapas descritas na figura 3:

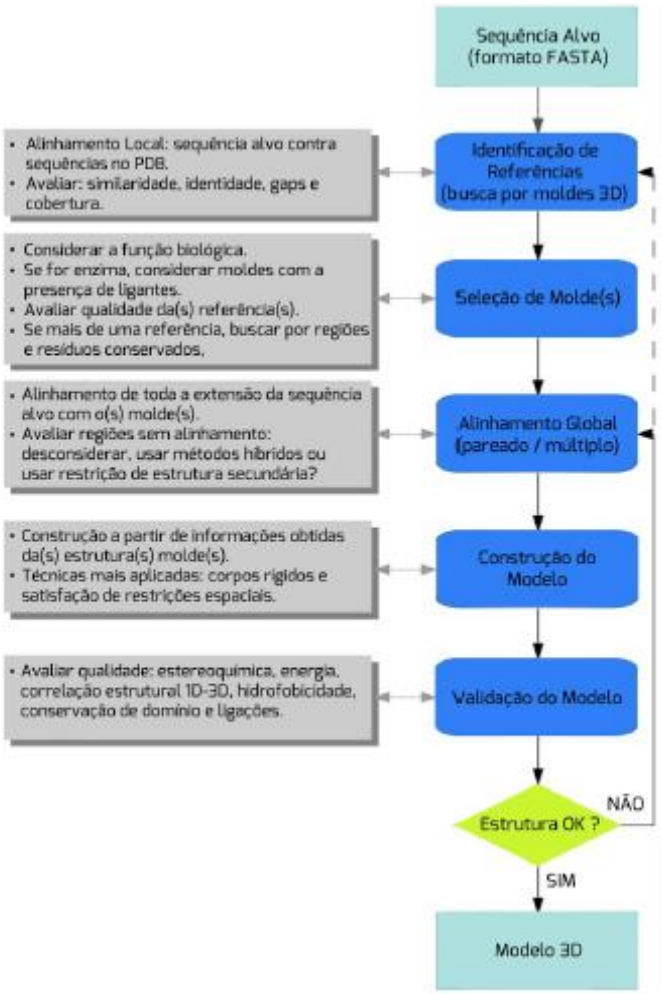


Figura 3. Cinco Etapas da Modelagem.

4.2 Identificação de referências

Tem por objetivo identificar sequências de aminoácidos de proteínas resolvidas experimentalmente que possuam similaridade com a sequência da proteína de interesse

(sequência alvo), cujas estruturas serão empregadas posteriormente como moldes. Essa identificação pode ser feita através de algoritmos de alinhamento, sendo selecionadas como referências as proteínas que possuírem os maiores índices de similaridade e identidade (suficientes para se inferir homologia entre as sequências), menores índices de gaps e a maior cobertura da sequência (relação entre a quantidade de aminoácidos alinhados entre as duas sequências e o tamanho total da sequência alvo).

4.3 Seleção dos moldes

Dentre as referências, é necessário escolher uma ou mais estruturas que servirão de molde para a construção do modelo 3D final. Nesta etapa, é imprescindível a análise do papel biológico da proteína de interesse. Os critérios de seleção podem incluir: i) a proteína de interesse e o possível molde pertencem a uma mesma família de proteínas; ii) ambas desempenham preferencialmente a mesma função ou tenham funções correlacionadas; iii) as estruturas resolvidas experimentalmente possuam alta qualidade (por exemplo, resolução ≤ 2 Å, fator R < 20%); iv) em tratando-se de uma enzima, é recomendado o uso de um molde cuja estrutura já tenha sido resolvida experimentalmente com seu substrato, ligante ou modulador. Na escolha de mais de uma estrutura molde, é importante realizar o alinhamento estrutural entre estas de forma a identificar regiões conservadas, sítios de ligação, águas estruturais e ligações dissulfeto conservadas.

4.4 Alinhamento entre as sequências

Uma vez escolhida(s) a(s) estrutura(s) molde, é necessário realizar alinhamento entre as sequências alvo e molde de forma a garantir que toda a proteína de interesse seja modelada (agora empregando programas como Clustal, T-Coffee e Muscle). Um alinhamento com mais de 40% de identidade é o suficiente para gerar um modelo confiável. Entretanto, é importante lembrar que o modelo final será uma representação desse alinhamento gerado. Portanto, regiões sem alinhamento significativo com o molde são preditas tridimensionalmente (quando preditas) sem grande confiabilidade, usando geralmente dados estatísticos gerais sobre estruturas de proteínas. Para as regiões sem alinhamento, deve-se considerar: i) a posição dessa região na

sequência de aminoácidos, verificando-se possíveis sítios de clivagem (principalmente em porções N- e C-terminal); ii) o tamanho dessa porção, considerando-se a possibilidade de formação de um novo domínio até então não identificado nessa família; iii) se são porções transmembranares, sejam preditas *in silico* (por exemplo, através das ferramentas TMHMM, HMMTOP, TMPred) ou já descritas em literatura porém ausentes nas estruturas molde; iv) o tipo de estrutura secundária predita in silico por mais de uma ferramenta (tais como PSIPRED, PHYRE, JUFO e PORTER), usando as regiões de consenso entre elas como informação de restrição de tipo de estrutura secundária durante a etapa de construção do modelo. Alternativamente, métodos híbridos podem ser aplicados para a predição de porções sem alinhamento. Para essas regiões, aplicam- se os métodos de predição de enovelamento ou primeiros princípios e usa-se a melhor estrutura predita como mais um molde para o método de modelagem por homologia.

4.5 Construção do modelo

A partir do alinhamento global entre as sequências alvo e molde, via modelagem por homologia, algoritmos específicos irão transferir as informações extraídas da estrutura 3D da proteína molde para o modelo. As técnicas mais aplicadas são as de construção usando corpos rígidos e por satisfação de restrições espaciais. A técnica de construção usando corpos rígidos constrói um modelo por partes, baseando-se na conservação de estruturas entre proteínas homólogas ou com grau significativo de identidade. As regiões estruturalmente conservadas da proteína de interesse são definidas através de predição de estruturas secundárias. Essas regiões são alinhadas com o molde, considerando se a média das posições dos Cα das sequências de aminoácidos das regiões estruturalmente conservadas. As regiões que não satisfazem as exigências são chamadas de regiões variáveis. Essas compreendem, geralmente, porções de alças que conectam as regiões conservadas. A cadeia principal dessas regiões pode ser obtida em bancos de dados específicos de estruturas, que apresentam conjuntos de alças classificados pelo número de aminoácidos e pelo tipo de estruturas secundárias que conectam. Após a inserção das regiões de alças, um modelo inicial do esqueleto peptídico estará pronto, restando apenas a inserção das cadeias laterais dos aminoácidos através de busca em bibliotecas de rotâmeros. Como exemplo de programa baseado nesta técnica, pode-se mencionar o portal Swiss-Model.

A segunda técnica mais comum, a construção por satisfação de restrições espaciais, inicia-se pelo alinhamento entre as sequências alvo e molde, extraindo-se desse molde suas

restrições espaciais (distâncias e ângulos) e transferindo-as para o modelo. Por exemplo, o tamanho das ligações e seus ângulos preferenciais são obtidos de campos de força. Dessa forma, é possível limitar o número de possíveis conformações que o modelo pode assumir. A principal característica dessa técnica é a obtenção empírica das restrições espaciais, expressas por funções de probabilidade, a partir de bancos de dados contendo informações sobre alinhamentos entre estruturas proteicas de alta resolução. As restrições espaciais e os termos de energia são combinados em uma função objetivo, sendo submetida a métodos de otimização por gradiente conjugado e recozimento simulado, visando a minimização das violações das restrições espaciais. Como exemplo de emprego desta técnica, pode-se citar o programa Modeller.

4.6 Validação do modelo

Após a construção do modelo, é necessário identificar possíveis erros relacionados aos métodos empregados, à escolha das referências e ao alinhamento entre as sequências alvo e molde. Caso o modelo seja caracterizado como de má qualidade, todo o protocolo anterior deve ser revisto no intuito de se melhorar o alinhamento, escolher outros moldes ou até mesmo decidir-se pelo uso de outros métodos. Por ser dependente de uma estrutura 3D resolvida experimentalmente, a técnica de modelagem por homologia possui certas limitações, tais como: i) nem sempre se consegue uma estrutura molde para a proteína de interesse; ii) o grau de similaridade conseguido entre as sequências alvo e molde pode ser pequeno (<30% de identidade), mesmo em regiões do sítio ativo, inviabilizando o emprego desta técnica; iii) por vezes, as sequências que podem servir como moldes possuem qualidade insuficiente para a construção de um modelo adequado. Nesses casos, como citado anteriormente, o uso adicional de informações, como a identificação de regiões transmembranares, a predição de regiões de peptídeo sinal, a predição de tipo de estrutura secundária, a predição do tipo de enovelamento e a verificação da existência de dados teóricos e experimentais quanto à existência, quantidade e localização de porções transmembranares, ligantes e número e tipo de cadeias podem contribuir tanto na construção de modelos tridimensionais como na anotação funcional de sequências. No caso de análises em larga escala de conjuntos de proteínas, e até mesmo de genomas inteiros, todo esse processo deve ser realizado para cada proteína de interesse. Considerando o tempo gasto em cada uma dessas etapas, é interessante o uso de métodos automatizados que podem ser empregados como um filtro inicial para a detecção de quais proteínas podem ser modeladas por modelagem por homologia e para a obtenção de um modelo inicial para cada uma dessas

proteínas, a ser otimizado individualmente. Como exemplo de programa usado para a análise em larga escala de sequências de proteínas, citamos o programa MHOLline. 7.6.

4.7 Predição do enovelamento

O método de predição do enovelamento ou threading parte da ideia de observações de que a estrutura 3D é mais conservada que a sequência, de forma que mesmo sequências com pouca similaridade podem possuir estruturas muito semelhantes, o que limita o número de enovelamentos que proteínas podem assumir. Atualmente, mais de 1.000 tipos de enovelamento já foram registrados, e acredita-se que esse valor não ultrapasse a previsão máxima de 7.000 tipos. Nesse método, também são usadas proteínas com estruturas 3D conhecidas e depositadas no PDB, de onde as informações sobre os tipos de enovelamento são extraídas e armazenadas em bancos de dados de tipos de enovelamentos. Como exemplo, citamos o CATH (Class, Architecture, Topology, Homology) e o SCOP (Structural Classification of Proteins). O método de predição do enovelamento é assim menos dependente da proximidade evolutiva entre a sequência de aminoácidos da proteína de interesse e seus possíveis moldes, ou seja, as sequências podem apresentar baixa identidade. O método é portanto aplicável quando o alinhamento entre a estrutura 1ária da proteína de interesse e de uma ou mais proteínas de referência (moldes) apresentam uma identidade entre 20% e 30%. No problema de PSP via predição do enovelamento tenta-se ajustar a estrutura 1ária da proteína de interesse aos tipos de enovelamentos de proteínas conhecidos, analisando principalmente as conservações de estruturas 2árias. Esse método pode ser dividido nas seguintes etapas: i) Reconhecimento do tipo de enovelamento pela análise das principais propriedades da proteína de interesse (tais como estrutura secundária, polaridade de cadeias laterais e hidrofobicidade); ii) Construção do melhor alinhamento possível entre a sequência de aminoácidos da proteína de interesse e estruturas depositadas em bancos de dados. Alguns métodos baseiam-se na construção de modelos simplificados (como modelos baseados em Cα) da proteína de interesse a partir da estrutura 3D de possíveis moldes, e avaliam a qualidade do modelo através da otimização de funções objetivo (geralmente não-lineares). Essas funções podem considerar, por exemplo, resultados de alinhamentos múltiplos de sequências e de estruturas secundárias, matrizes de substituição para cada aminoácido dentro de uma família específica de proteínas e penalização de gaps; iii) Escolha do(s) melhor(es) molde(s) para a construção da estrutura 3D da proteína de interesse, geralmente baseada em funções de predição de erro/qualidade entre os possíveis modelos

simplificados e seu(s) molde(s) (por exemplo, a função TM-score). A escolha dos melhores moldes por vezes é baseada em bibliotecas de fragmentos; iv) Construção do modelo 3D através de técnicas similares às empregadas na modelagem comparativa, por vezes valendo-se de ferramentas acopladas aos programas Swiss-Model ou Modeller. Alguns programas empregam, para as regiões sem molde, métodos por primeiros princípios. Como exemplo de programas para PSP via predição do enovelamento pode-se citar os programas HH-Pred e I-TASSER. As limitações dos métodos de predição do enovelamento vêm de dois pontos principais. O primeiro é similar ao observado para a modelagem comparativa, isto é, se a identidade entre a sequência alvo e as proteínas utilizadas na construção do banco de enovelamentos for muito baixa, é possível que o enovelamento daquela sequência simplesmente não esteja representado no banco. Assim, o método pode construir um modelo completamente errado. A outra limitação é que os modelos apresentam uma resolução relativamente baixa, dificultando seu uso em estudos que exigem posicionamento preciso dos átomos como no caso do atracamento, como veremos.

4.7 Análise de qualidade

A qualidade de um modelo é determinada por um conjunto de fatores, tais como comprimentos de ligação, planaridade das ligações peptídicas, planaridade dos anéis e ângulos de torção nas cadeias principal (ou seja, esqueleto peptídico) e laterais, quiralidade, impedimento estérico, energia e funcional. Adicionalmente, nos métodos baseados no uso de estruturas moldes resolvidas experimentalmente, para um modelo ser considerado de boa qualidade é recomendado que o valor de RMSD obtido pela sobreposição da cadeia peptídica de regiões conservadas do modelo gerado e da estrutura molde esteja entre 1 Å e 2 Å. Dentre as análises a serem feitas, recomenda-se as seguintes: i) Estereoquímica: consiste em analisar os aspectos tridimensionais de uma molécula, a fim de se verificar a estabilidade conformacional da mesma. Nesta análise, são detectadas regiões de tensão angular e torcional, impedimentos estéricos e quiralidades. Além destes, com a análise do gráfico de Ramachandran é possível identificar, através da correlação entre os ângulos φ e ψ, quais resíduos encontram-se fora das regiões energeticamente favoráveis, possibilitando uma melhora no modelo final. Exemplos de programas que realizam estas análises incluem os programas Procheck e Molprobity. ii) Energia: são métodos baseados em minimização de funções de energia. A análise dos valores normalizados da função (como o DOPE normalizado do Modeller) ajuda a avaliar (ao menos estatisticamente) quão próximo o modelo gerado está de proteínas que possuem um mesmo perfil molecular ou até o mesmo tipo

de enovelamento. Esses métodos podem considerar a relação entre a estrutura 1D-3D, ponderar a propensão de cada aminoácido estar em um tipo de estrutura secudária, a probabilidade de dois resíduos estarem em contato e até mesmo o tipo de função que a proteína desempenha. Alguns programas bastante usados para estas análises incluem Verify3D, ProSa, QMEAN e PROVE. iii) Funcional: envolve a comparação do modelo obtido com aspectos funcionais ou mesmo estruturais (sem resolução atomística) determinados por métodos experimentais. Por exemplo, diversas famílias de proteínas possuem resíduos específicos associados à função (como a tríade catalítica em serino proteases ou resíduos ligadores de metais em metaloproteínas). Assim, o modelo gerado deve apresentar tais resíduos nas suas localizações específicas para explicar dados experimentais prévios. Ainda, métodos como dicroismo circular, infravermelho e RMN podem oferecer informações importantes sobre o estado conformacional da proteína em meio biológico, validando o modelo obtido. Mesmo que as estratégias de análise anteriores indiquem um modelo de elevada qualidade, se o mesmo não for capaz de apresentar ou explicar características conhecidas previamente, não poderá ser considerado totalmente válido. Durante o CASP a análise de qualidade dos modelos assume um caráter diferente, uma vez que os avaliadores conhecem a estrutura nativa. Nesse caso, a métrica empregada para comparar a estrutura nativa com os modelos gerados pelos diferentes métodos é o Global Distance Test – GDT. Trata-se de uma medida potencialmente mais acurada, uma vez que é menos sensível a discrepâncias muito grandes, oriundas de regiões de voltas que são naturalmente flexíveis.

4.8 Refinamento do modelo

Após a análise do modelo, caso a qualidade não tenha sido satisfatória, algumas estratégias de refinamento no melhor modelo obtido podem ser suficientes para a obtenção de um modelo final de boa qualidade. Dentre os principais tipos de refinamento podemos citar: i) Local: através da análise estereoquímica pode-se identificar qual resíduo está violando seus valores limites dentro de sua vizinhança, o que geralmente é resolvido com o reposicionamento de sua cadeia lateral. Em alguns casos, é necessário realizar etapas de otimização somente de regiões de alças, principalmente de regiões ricas em glicina. É sempre importante observar violações causadas por prolinas nas extremidades de regiões de estruturas em hélice ou folha. ii) Imposição de restrições: após a análise de resultados de métodos de predição de estrutura secundária, pode-se verificar no modelo gerado quais regiões não possuem ou possuem uma baixa similaridade de sequência com o(s) molde(s) usado(s), ou não obedecem ao tipo correto de estrutura secundária

predita. Para corrigir isso, é necessário refazer o modelo 3D impondo ao algoritmo de construção o uso de restrições de tipo de estrutura secundária para essas regiões. iii) Dinâmica molecular: Os métodos de simulação por dinâmica molecular têm sido empregados na melhora de modelos gerados tanto por técnicas baseadas em modelagem comparativa quanto por primeiros princípios. Simulações em solvente explícito ajudam a acomodar a estrutura 3D do modelo melhorando, principalmente, os ângulos φ e ψ de resíduos em regiões desfavoráveis no gráfico de Ramachandran. O tempo de simulação é variável de acordo com a complexidade do sistema e com o grau de refinamento que se deseja obter. É importante destacar que simulações por dinâmica molecular para estruturas transmembranares, apesar de bastante recomendado, necessitam especial atenção, pois se deve considerar o modelo de membrana a ser empregado, a forma de inserção do modelo 3D da proteína na membrana e o tempo de equilibração do sistema costuma ser maior que em proteínas simuladas apenas em solvente.

4.9 Aplicações de modelos

A aplicabilidade de um modelo 3D está diretamente relacionada com a acurácia com que este foi gerado. Esta acurácia pode ser avaliada pelo grau de similaridade entre as estruturas 3D da proteína predita e da proteína molde, através do cálculo do desvio médio quadrático (RMSD), que mede as distâncias interatômicas. De acordo com sua acurácia, os modelos 3D gerados por métodos teóricos podem ser aplicados em: i) Estudos de predição funcional e busca por novos alvos moleculares em organismos patogênicos; ii) Planejamento racional de fármacos baseado na estrutura do receptor biológico; iii) Estudos de variação conformacional por dinâmica molecular; iv) Planejamento de experimentos de mutagênese sítio-dirigida, fornecendo informações sobre possíveis mutações para testar hipóteses funcionais; v) Simulações de interações entre proteínas; vi) Auxiliar no refinamento de estruturas resolvidas por cristalografia de raios-X e por experimentos de RMN.

4. 10 Tutorial de Modelagem

Problematização (hipotética):

A bactéria *Vibrio cholerae*, causadora da doença conhecida como cólera adquiriu resistência aos principais antibióticos conhecidos. Portanto, é necessário identificar a estrutura de novos alvos moleculares de antibióticos nesta bactéria para o desenvolvimento de novos fármacos. Neste contexto, a enzima Alanina racemase é de vital importância para o metabolismo desta bactéria. Esta enzima catalisa a conversão da L-alanina em D-alanina, um aminoácido necessário para a síntese da parece celular da bactéria. Sem a D-alanina, a bactéria não consegue se multiplicar, o que impede a proliferação da infecção. A inibição desta enzima não afeta os mamíferos, pois nós não necessitamos de D-alanina para sobreviver.

Objetivo:

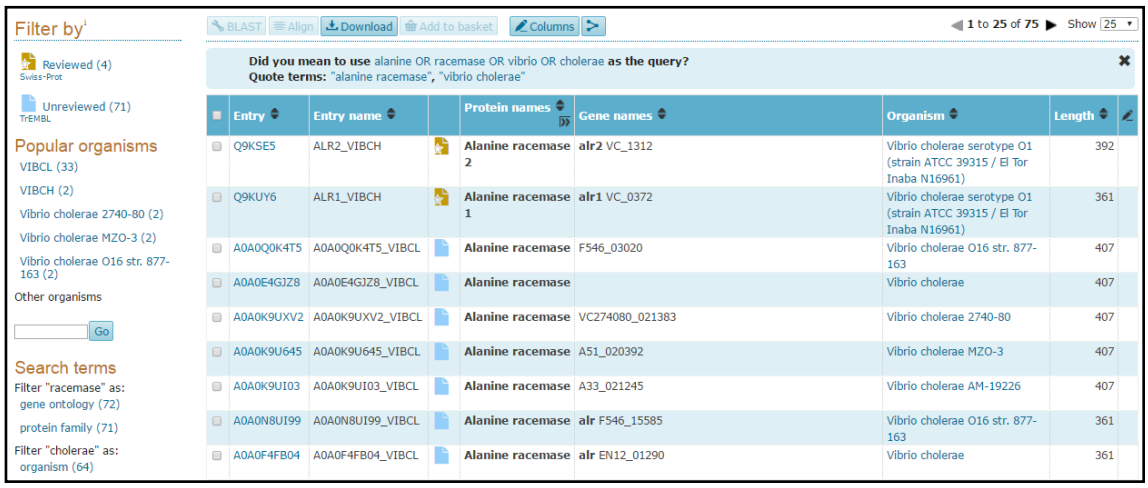
Modelar a estrutura 3D da enzima alanina racemase (E.C. 5.1.1.1) de *Vibrio cholerae* complexada com um ligante, de modo a ser utilizada como molde estrutural para o desenvolvimento de inibidores específicos (*drug desing*).

Passos do tutorial

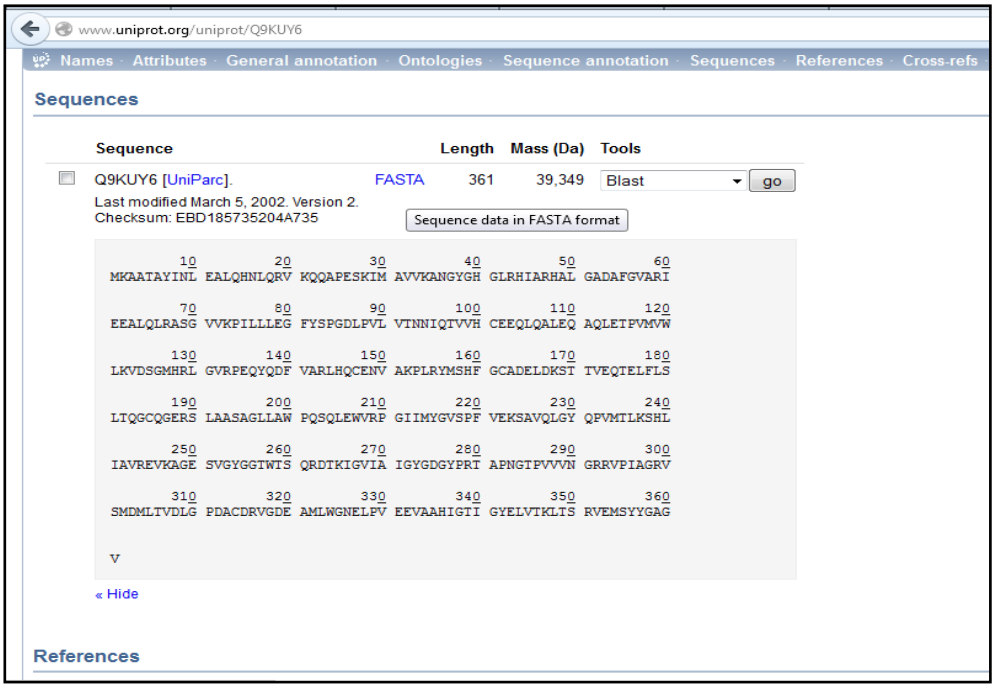
1) Criar uma pasta de trabalho no ambiente Linux: **modelagemAR**

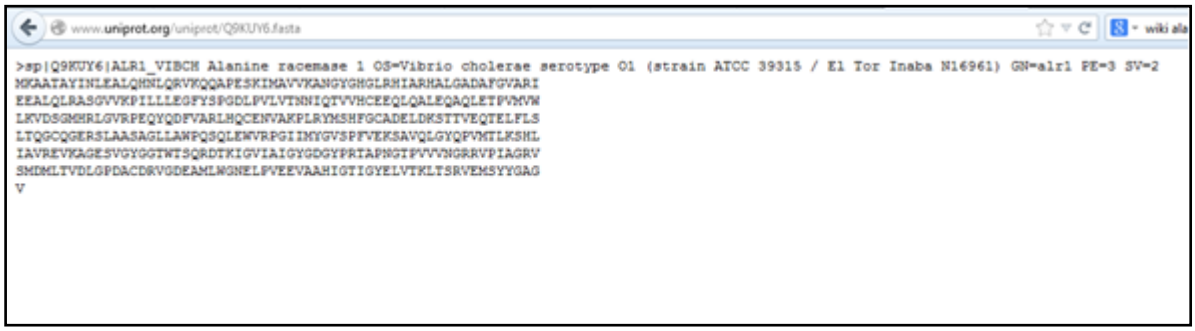
2) Buscar a seqüência de aminoácidos da enzima alvo no Uniprot (www.uniprot.org)

alanine racemase vibrio cholerae



Escolher a sequência **Q9KUY6** e baixar no formato fasta.



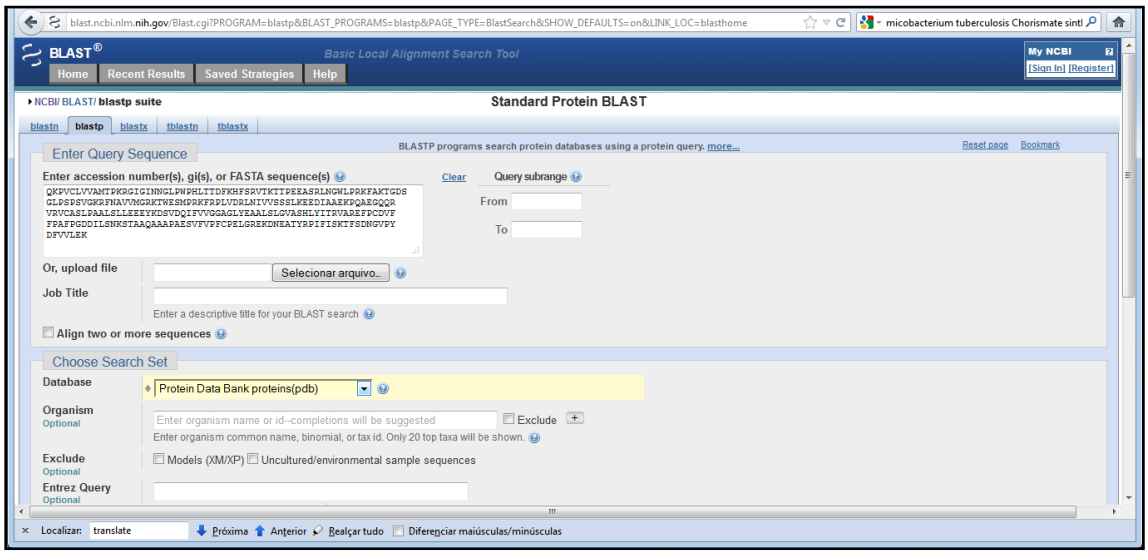


3) Selecionar, copiar e salvar a sequência em um arquivo na pasta de trabalho, como sequencia.txt

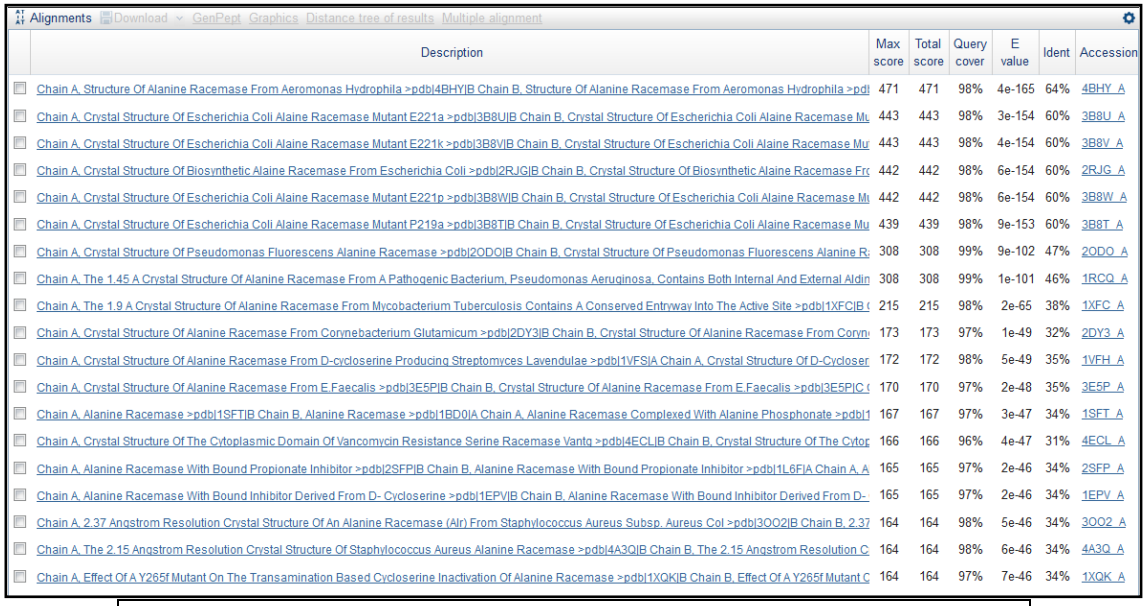
4) retirar a metionina (M) de início.

5) Procurar por moldes estruturais no NCBI BLASTp : http://blast.ncbi.nlm.nih.gov/Blast.cgi

- Colar a seqüência da *Vcholerae*AR e selecionar a base de dados **protein data bank (pdb)**



6) Escolher o molde com maior identidade e com um inibidor ligado.



Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
Chain A, Structure Of Alanine Racemase From Aeromonas Hydrophila >pdb|4BHY|B Chain B, Structure Of Alanine Racemase From Aeromonas Hydrophila >pdb	471	471	98%	4e-165	64%	4BHY_A
Chain A, Crystal Structure Of Escherichia Coli Alaine Racemase Mutant E221a >pdb|3B8U|B Chain B, Crystal Structure Of Escherichia Coli Alaine Racemase Mu	443	443	98%	3e-154	60%	3B8U_A
Chain A, Crystal Structure Of Escherichia Coli Alaine Racemase Mutant E221k >pdb|3B8V|B Chain B, Crystal Structure Of Escherichia Coli Alaine Racemase Mu	443	443	98%	4e-154	60%	3B8V_A
Chain A, Crystal Structure Of Biosynthetic Alaine Racemase From Escherichia Coli >pdb|2RJG|B Chain B, Crystal Structure Of Biosynthetic Alaine Racemase Fr	442	442	98%	6e-154	60%	2RJG_A
Chain A, Crystal Structure Of Escherichia Coli Alaine Racemase Mutant E221p >pdb|3B8W|B Chain B, Crystal Structure Of Escherichia Coli Alaine Racemase Mu	442	442	98%	6e-154	60%	3B8W_A
Chain A, Crystal Structure Of Escherichia Coli Alaine Racemase Mutant P219a >pdb|3B8T|B Chain B, Crystal Structure Of Escherichia Coli Alaine Racemase Mu	439	439	98%	9e-153	60%	3B8T_A
Chain A, Crystal Structure Of Pseudomonas Fluorescens Alanine Racemase >pdb|2ODO|B Chain B, Crystal Structure Of Pseudomonas Fluorescens Alanine R	308	308	99%	9e-102	47%	2ODO_A
Chain A, The 1.45 A Crystal Structure Of Alanine Racemase From A Pathogenic Bacterium, Pseudomonas Aeruginosa, Contains Both Internal And External Aldin	308	308	99%	1e-101	46%	1RCQ_A
Chain A, The 1.9 A Crystal Structure Of Alanine Racemase From Mycobacterium Tuberculosis Contains A Conserved Entryway Into The Active Site >pdb|1XFC|B	215	215	98%	2e-65	38%	1XFC_A
Chain A, Crystal Structure Of Alanine Racemase From Corynebacterium Glutamicum >pdb|2DY3|B Chain B, Crystal Structure Of Alanine Racemase From Coryn	173	173	97%	1e-49	32%	2DY3_A
Chain A, Crystal Structure Of Alanine Racemase From D-cycloserine Producing Streptomyces Lavendulae >pdb|1VFS|A Chain A, Crystal Structure Of D-Cycloser	172	172	98%	5e-49	35%	1VFH_A
Chain A, Crystal Structure Of Alanine Racemase From E.Faecalis >pdb|3E5P|B Chain B, Crystal Structure Of Alanine Racemase From E.Faecalis >pdb|3E5P|C	170	170	97%	2e-48	35%	3E5P_A
Chain A, Alanine Racemase >pdb|1SFT|B Chain B, Alanine Racemase >pdb|1BD0|A Chain A, Alanine Racemase Complexed With Alanine Phosphonate >pdb|1	167	167	97%	3e-47	34%	1SFT_A
Chain A, Crystal Structure Of The Cytoplasmic Domain Of Vancomycin Resistance Serine Racemase VanTg >pdb|4ECL|B Chain B, Crystal Structure Of The Cytop	166	166	96%	4e-47	31%	4ECL_A
Chain A, Alanine Racemase With Bound Propionate Inhibitor >pdb|2SFP|B Chain B, Alanine Racemase With Bound Propionate Inhibitor >pdb|1L6F|A Chain A, A	165	165	97%	2e-46	34%	2SFP_A
Chain A, Alanine Racemase With Bound Inhibitor Derived From D- Cycloserine >pdb|1EPV|B Chain B, Alanine Racemase With Bound Inhibitor Derived From D-	165	165	97%	2e-46	34%	1EPV_A
Chain A, 2.37 Angstrom Resolution Crystal Structure Of An Alanine Racemase (Alr) From Staphylococcus Aureus Subsp. Aureus Col >pdb|3OO2|B Chain B, 2.37	164	164	98%	5e-46	34%	3OO2_A
Chain A, The 2.15 Angstrom Resolution Crystal Structure Of Staphylococcus Aureus Alanine Racemase >pdb|4A3Q|B Chain B, The 2.15 Angstrom Resolution C	164	164	98%	6e-46	34%	4A3Q_A
Chain A, Effect Of A Y265f Mutant On The Transamination Based Cycloserine Inactivation Of Alanine Racemase >pdb|1XQK|B Chain B, Effect Of A Y265f Mutant C	164	164	97%	7e-46	34%	1XQK_A

Escolher: **2RJG -> 2RJH**

7) Baixar do PDB (www.pdb.org) a estrutura do molde e visualizá-la no programa **coot** .

8) Editar o arquivo do molde (2RJH.pdb) de modo a deixar somente as cadeias A e B (homodímero, unidade biológica) mais o ligante. Apagar todas as águas e renumerar os ligantes na seqüência.

9) Baixar os scripts de execução, a partir do tutorial do programa modeller.

(http://salilab.org/modeller/tutorial/) Ou localmente: (usr/lib/modeller9.12/examples/)

Para este exercício, usar os scripts editados.

10) Sintaxe de execução no terminal: ***mod9.12 script.py***

11) Para executar este tutorial, siga os seguintes passos:

1 – Alinhar a seqüência do molde com a do modelo (script aling2d.py)

2 – Conferir o output do alinhamento *.pap *.ali

3 – Gerar 10 modelos (*para publicação, gerar acima de 100 modelos*) e listar os 5 melhores (script model-single.py)

12) Validação dos melhores modelos:

- Crie uma subpasta chamada **procheck**

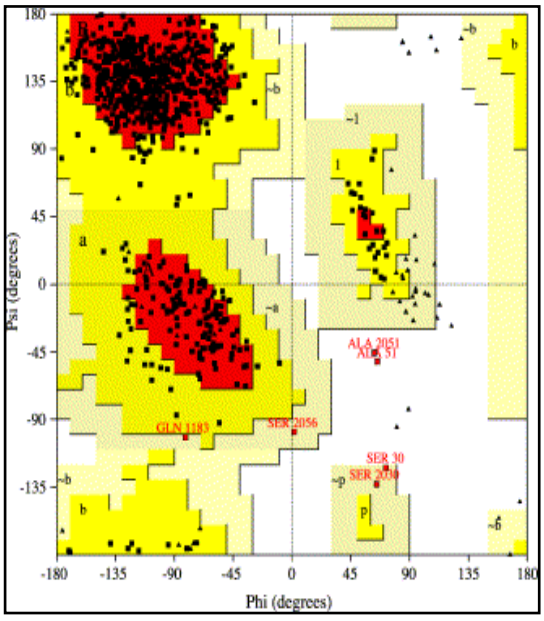
- Copie os 5 melhores modelos para a pasta procheck.

- Gerar o gráfico de Ramachandram para cada modelo usando a seguinte sintaxe:

procheck modelo01.pdb 2.0 (enter)

13) Verificar a qualidade do modelo pelo **gráfico de Ramachandram**

O gráfico de Ramachandran é particularmente útil porque ele define os resíduos que se encontram nas regiões energicamente mais favoráveis e desfavoráveis e orienta a avaliação da qualidade de modelos teóricos ou experimentais de proteínas.



Bons modelos tem > 90% dos resíduos em regiões permitidas.

5 Docking Molecular e Varredura Virtual

Trechos retirados do livro: ***Bioinformática: Da biologia à flexibilidade molecular****. Porto Alegre: e-book. Disponível em: <http://www.ufrgs.br/bioinfo/ebook/>.*

5.1 Introdução

Atracamento molecular (do inglês *molecular docking*) é uma técnica computacional que visa prever o modo de ligação e dos detalhes do reconhecimento molecular entre duas macromoléculas. Os métodos de atracamento molecular envolvem desafios teórico-computacionais formidáveis, e se dividem em duas classes de métodos distintos: receptor-ligante e proteína-ligante.

Embora proteínas sejam os receptores mais comuns, outras biomoléculas também podem exercer este papel. Diversos fármacos, por exemplo, modulam diretamente o DNA que, assim, passa a ser o receptor alvo. Adicionalmente, fármacos podem atuar modificando propriedades físico-química da célula, sem necessariamente envolver um processo de atracamento, como na modulação da fluidez de membranas plasmáticas. Neste estudo, será dada mais ênfase aos métodos de atracamento proteína-ligante, contextualizados dentro da área de planejamento racional de fármacos baseado em estruturas.

As metodologias computacionais de atracamento proteína-ligante estão baseadas no modelo chave-fechadura, proposto por Emil Fischer em 1894. Neste modelo, o receptor proteico é associado à uma “fechadura”, e seu sítio de ligação ou sítio receptor é considerado como o “buraco da fechadura”. A possível “chave da fechadura” é o ligante, e a interação entre o ligante e a proteína está relacionada a uma das possíveis ações de “abrir ou trancar” a porta. O modelo chave-fechadura, contudo, induz a uma interpretação de que a “fechadura”, representada pela molécula receptora, é rígida. Entretanto, no meio biológico, tanto o ligante quanto a proteína são flexíveis, podendo modificar a sua conformação durante o processo de formação do complexo receptor- ligante. Uma visão mais adequada deste processo é denominada de encaixe induzido, onde tanto o ligante quanto a proteína se adaptam um ao outro durante o processo de reconhecimento molecular. De fato, a flexibilidade de uma proteína está diretamente associada à sua atividade, seja na catálise de reações enzimáticas, na transdução de sinais, no transporte através de proteínas de membrana, ou em mudanças conformacionais associadas a formas ativas e não ativas de proteínas.

O reconhecimento molecular proteína-ligante está baseado na complementaridade de características físico-químicas e estruturais das moléculas interagentes. As características físico-químicas definem o grau de afinidade e de especificidade do ligante pela proteína, e estão

relacionadas com as interações intermoleculares existentes no complexo. Estas interações incluem as ligações de hidrogênio, as interações provenientes do efeito hidrofóbico, as interações de van der Waals, as interações eletrostáticas e as ligações covalentes que possam ser formadas durante o processo de interação receptor-ligante. As características estruturais, por sua vez, estão associadas aos arranjos espaciais moleculares, dados por variações na orientação, posicionamento espacial e rotações de ligações químicas das moléculas interagentes. Ligantes e proteínas que possuem uma alta afinidade um pelo outro exibem as seguintes características: *i*) alto nível de complementaridade estérica, ou seja, a proteína e o ligante possuem uma alta porcentagem de suas superfícies de contato moleculares, definidas pelos raios de van der Waals atômicos, em contato próximo; *ii*) alta complementaridade de propriedades associadas às superfícies de contato moleculares (esta complementaridade pode ser tanto eletrostática, onde grupos polares/carregados do ligante ficam perto de grupos da proteína com polaridade/carga complementar, quanto relacionada à complementaridade de regiões hidrofóbicas); *iii*) o ligante geralmente se liga em uma conformação energeticamente favorável, e *iv*) interações repulsivas entre ligante e proteínas são minimizadas.

5.2 Interações proteína-ligante

Os principais tipos de interações intermoleculares envolvidas no reconhecimento molecular proteína-ligante incluem: *i*) ligações de hidrogênio; *ii*) interações de van der Waals; *iii*) interações iônicas; *iv*) interações hidrofóbicas; *v*) interações do tipo cátion-π; *vi*) interações envolvendo anéis aromáticos do tipo π-π e empilhamento-T, e *vii*) coordenação com íons metálicos. O efeito hidrofóbico origina-se do fato de que partes apolares do ligante e do sítio ativo interagem com o solvente, sendo que estas se encontram solvatadas por camadas de moléculas de água mais organizadas. A aproximação destas partes apolares, durante a interação proteína-ligante, liberam e desorganizam as moléculas de água, aumentando a entropia do sistema e consequentemente favorecem a formação do complexo proteína-ligante. O aumento na entropia do solvente associado ao ocultamento das superfícies apolares é chamado de efeito hidrofóbico. Este efeito destaca o papel fundamental do solvente aquoso no processo de reconhecimento molecular proteína-ligante. Em algumas situações, as moléculas de água assumem tal importância que sua presença é considerada estrutural, sendo por isso denominadas moléculas de água estruturais. Estas moléculas estão ligadas fortemente ao sítio ativo, e geralmente são conservadas em sítios de ligação de proteínas homólogas. A presença destas

moléculas nos sítios receptores de proteínas podem interferir no acesso do ligante ao sítio ativo e modificar o perfil de formação de ligações de hidrogênio, contribuindo portanto diretamente no sucesso das metodologias de atracamento proteína-ligante.

O processo de atracamento pode ser dividido em duas etapas principais:

i) **Função de busca** - investigação e predição da conformação e orientação de um ligante em seu sítio de ligação

ii) **Função de ranqueamento** - predição da afinidade em um complexo receptor-ligante, isto é, a energia livre de ligação (normalmente chamado na literatura de função *scoring*)

Essas funções são visadas para varreduras de bibliotecas imensas de compostos, permitindo a predição de estruturas mais aptas a se ligarem no sítio de determinado alvo.

5. 3 Tutorial Prático sobre Docking e Varredura Virtual – Windows XP

Validação protocolo de *docking*: Método do Redocking

Princípio: Se uma proteína foi cristalizada ou Modelada e minimizada na presença de um ligante, admite-se que a posição (pose) do ligante na estrutura final de menor energia seja a mais estável e, portanto, a pose mais provável encontrada na proteína em solução. Portanto, ao fazer o redocking do ligante na estrutura da proteína que foi cristalizada/modelada na presença deste mesmo ligante, o programa de docagem deve ser capaz de "encontrar/reproduzir" esta pose de menor energia.

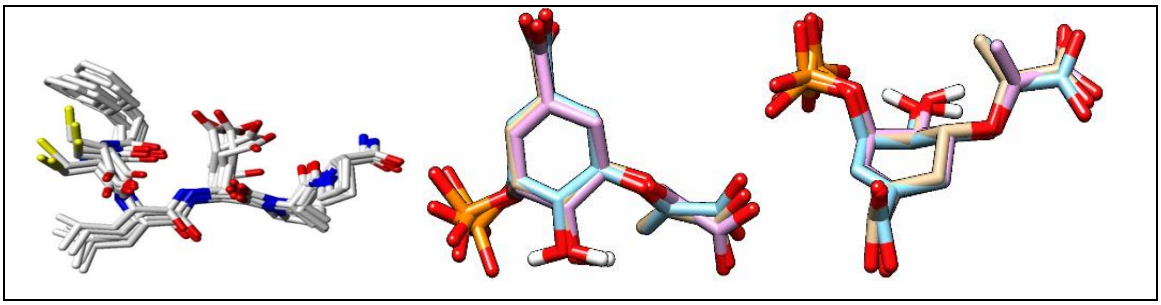
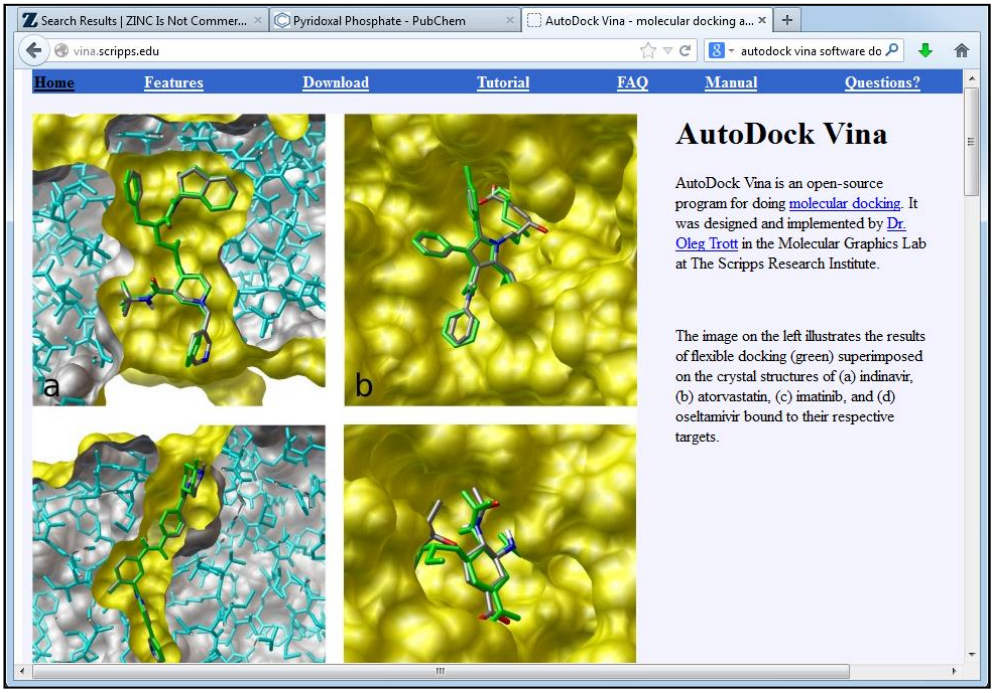


Figura 4. Redocking.

Programas utilizados neste tutorial

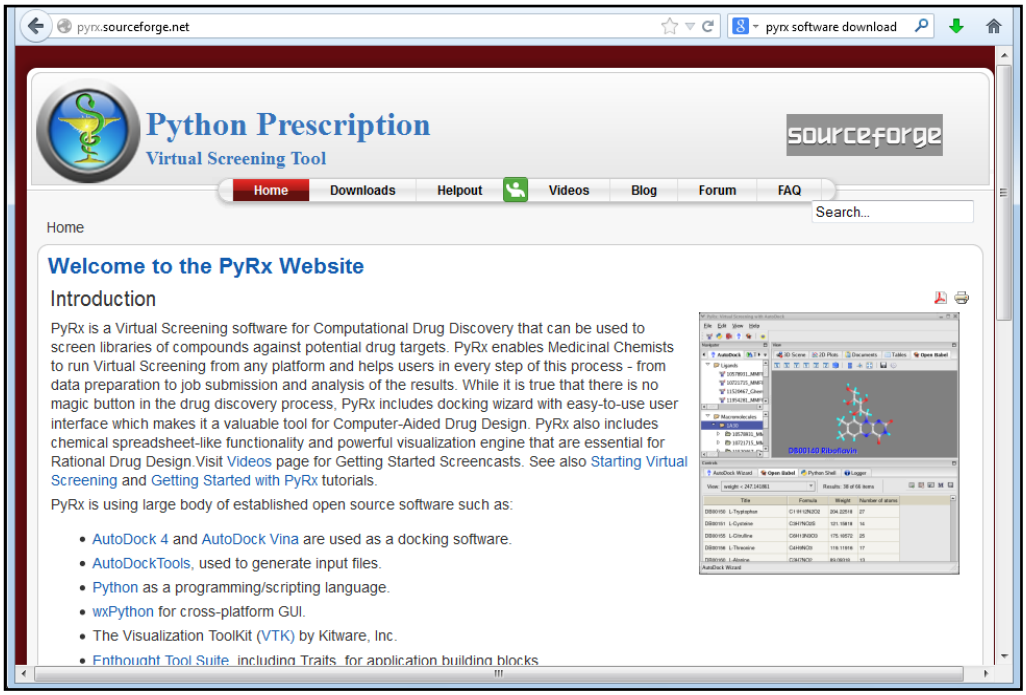
Para docagem**: Vina**

- Licença acadêmica gratuita
- http://vina.scripps.edu/



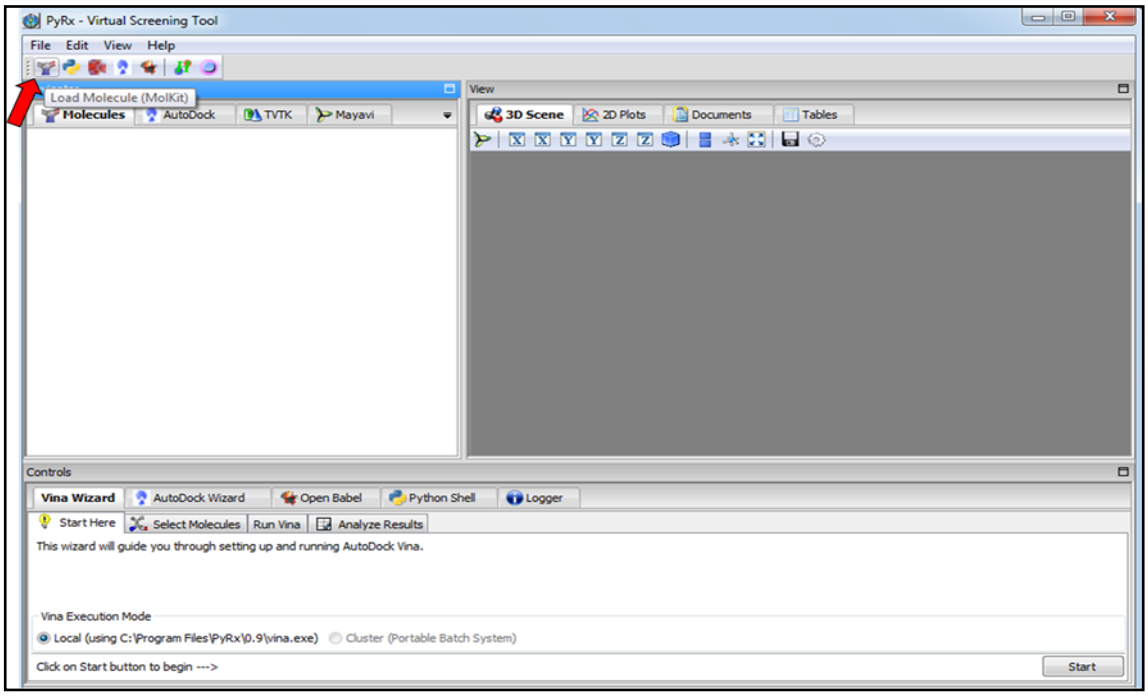
Interface gráfica: **Pyrx**

- Acompanha o programa Vina
- Também funciona para o programa Autodock
- Licença acadêmica gratuita
- http://pyrx.sourceforge.net/



Passos a seguir no tutorial:

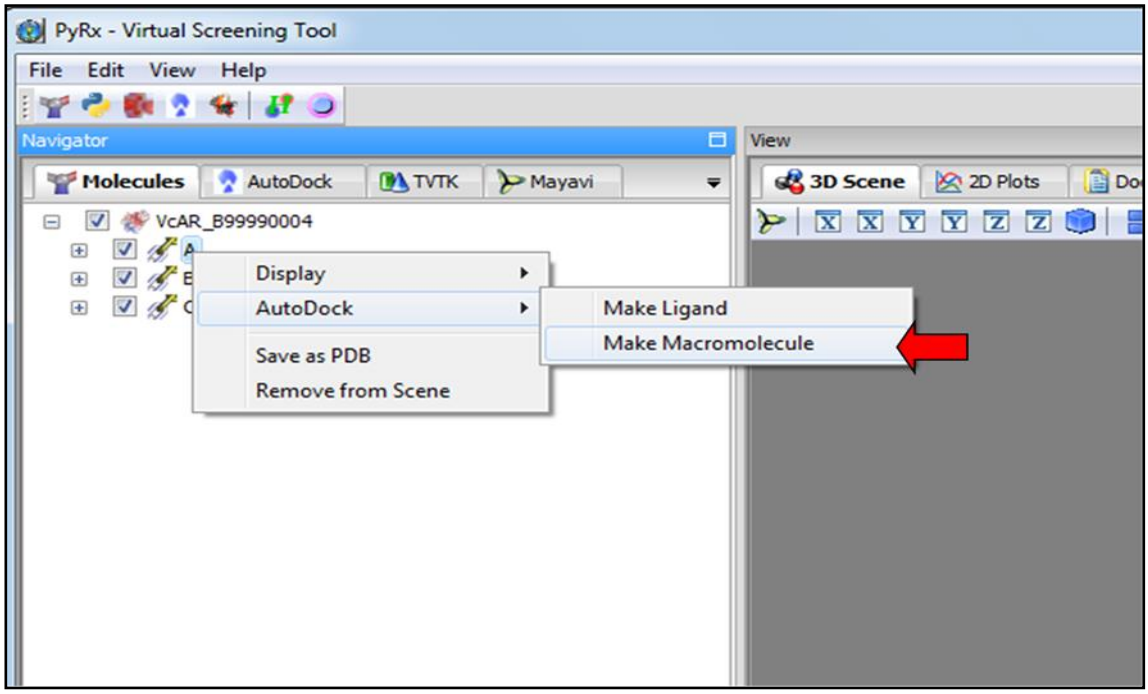
1) Carregar o programa, **File → Load Molecule**, depois, abra a estrutura modelada: **VcAR.B99990004.pdb**



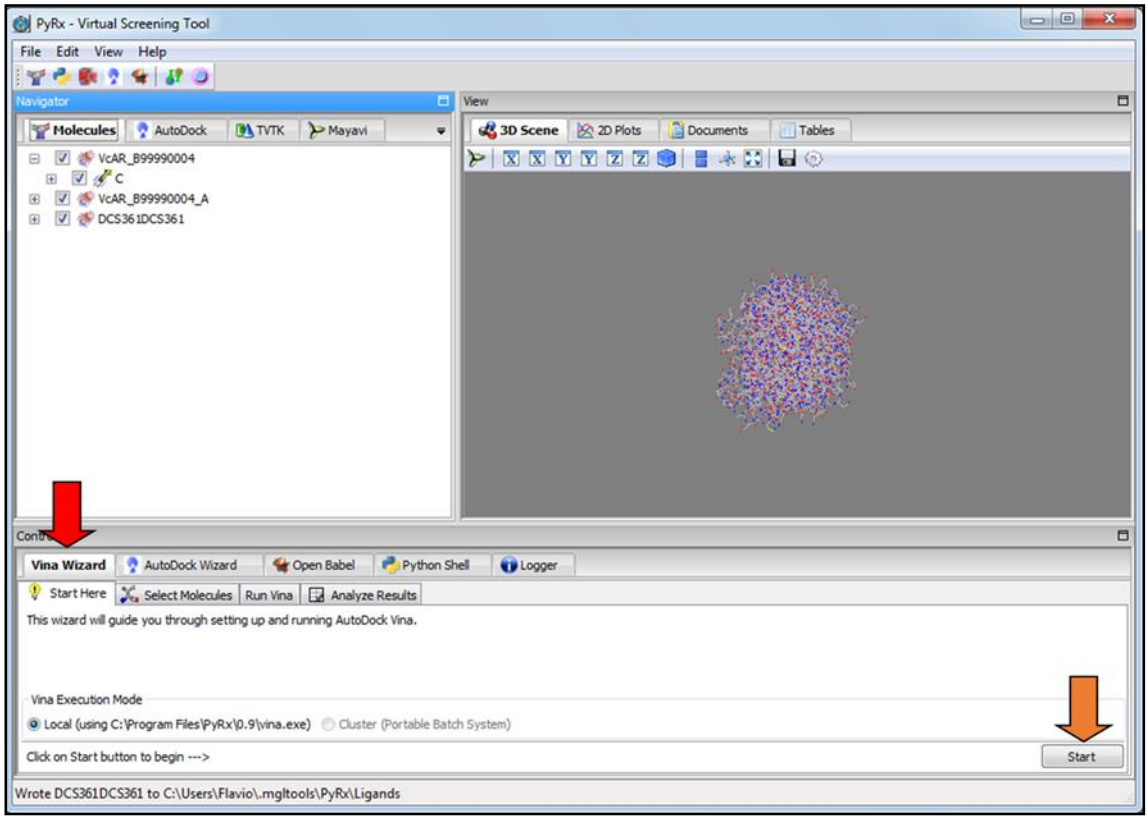
2) **Clicar com o botão direito do mouse sobre as letras:**

A → Autodock → Make Macromolecule

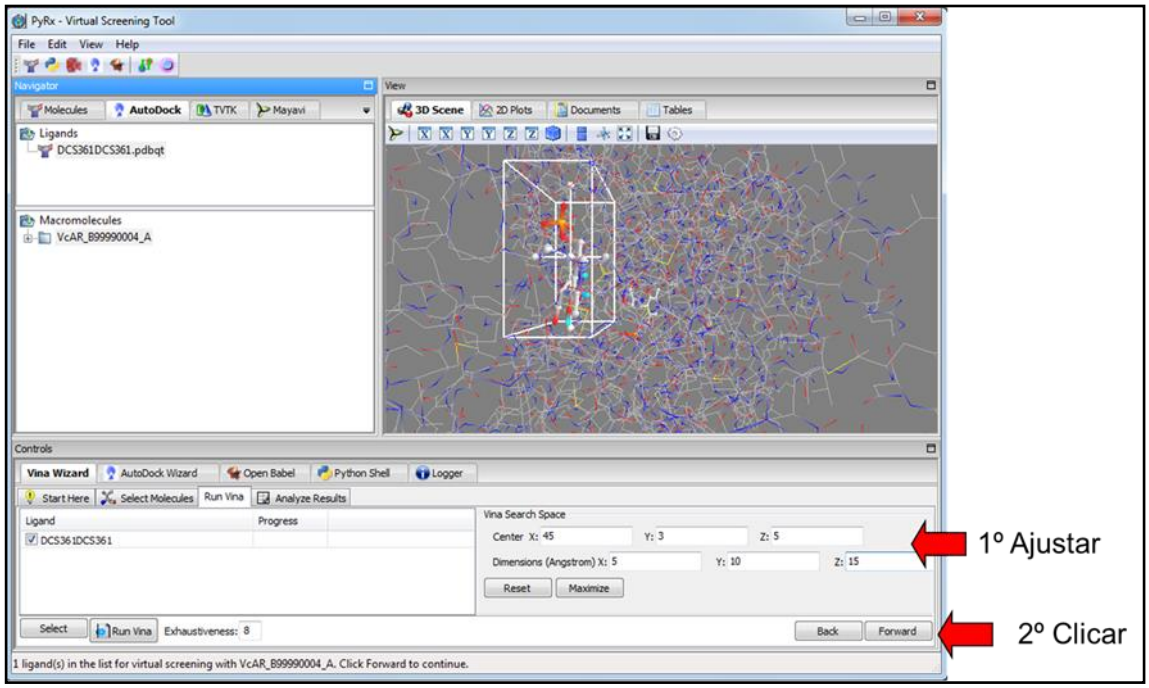
B → Autodock → Make Ligand



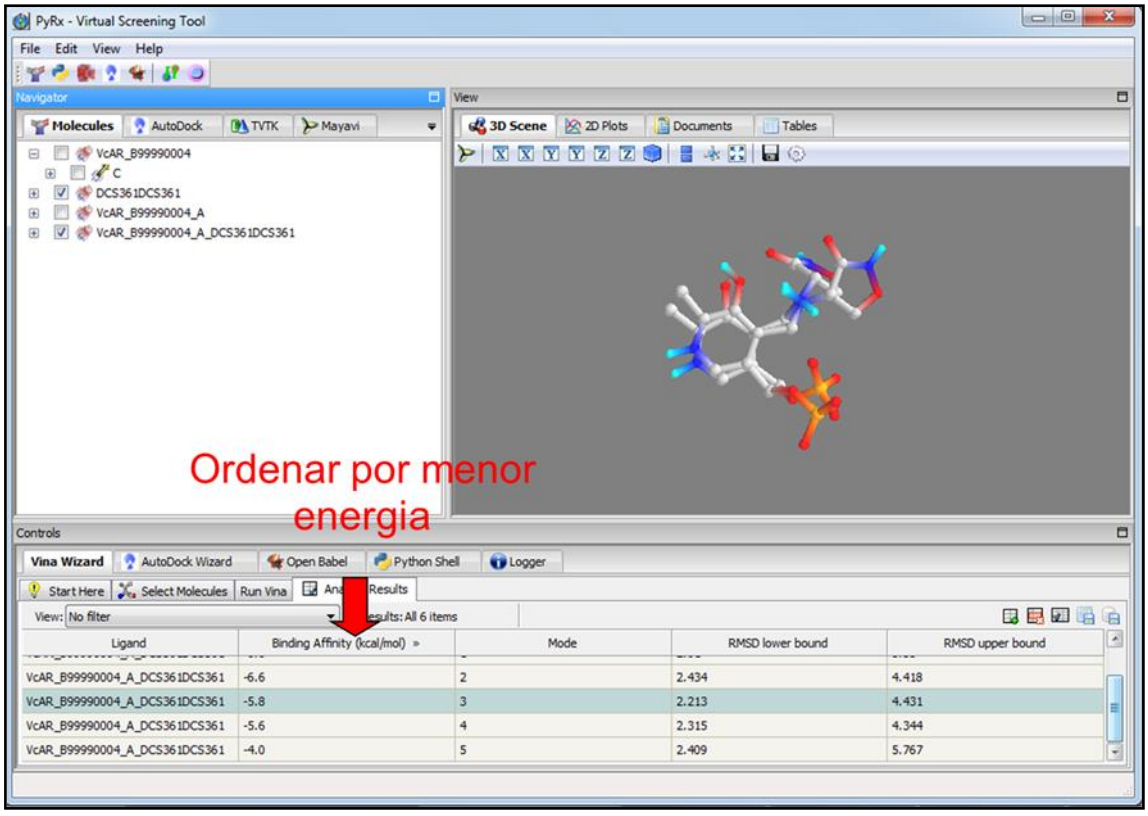
3) **Clicar na aba Vina Wizard e em seguida Start, depois Forward, até a próxima tela.**



4) **Parâmetros de busca: Centro do sítio ativo: 45, 3, 5**
Tamanho da caixa: 5, 10, 10



5) Selecionar a pose que melhor se sobrepõe ao ligante



6) Validação do protocolo

- ✓ Se o programa de docagem foi capaz de reproduzir a pose do ligante com um rmsd ~ 1,0 Å, o protocolo pode ser considerado validado e pode agora ser aplicado na busca de novos ligantes/inibidores para sua enzima.
- ✓ Caso isso não aconteça, você deve considerar modificar os parâmetros de busca e se, mesmo assim, não for possível reproduzir a pose modelada, você deve considerar mudar o programa/algoritmo de docagem.

5.4 Varredura virtual (virtual screening) usando o Vina na interface Pyrx

A varredura virtual (VS) consiste de duas etapas:

- ✓ **Primeira**: Montagem da biblioteca de ligantes a serem testados.
  - Ligantes com semelhança estrutural ao substrato natural.
  - Ligantes com semelhança estrutural aos inibidores já conhecidos.

- Ligantes de uma base de dados específica, exemplo: Catálogos de Produtos Naturais, Sigma-aldrich, Acros, MolPort, etc.

✓ **Segunda**: Uso de um programa e protocolo de docagem validados para sua estrutura.

- O programa vai testar cada um dos ligantes no sítio ativo especificado de seu alvo molecular e ranquear os melhores candidatos.
- O programa faz dois cálculos distintos, primeiro leva em consideração as várias poses possíveis em função dos pontos de rotação da molécula (search). O segundo cálculo leva em consideração as interações moleculares do ligante com seu alvo (pontes H, contatos de vdW, pontes salinas, etc) e ranqueia (ranking) cada uma das poses.

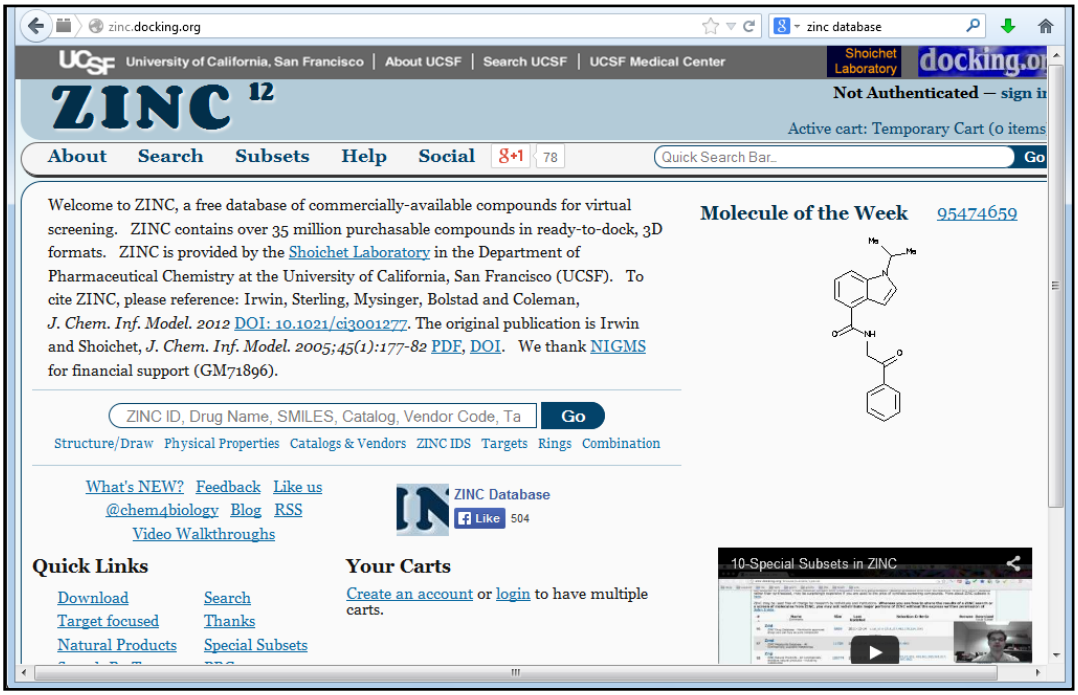
Principais bancos de dados de ligantes virtuais (gratuitos)

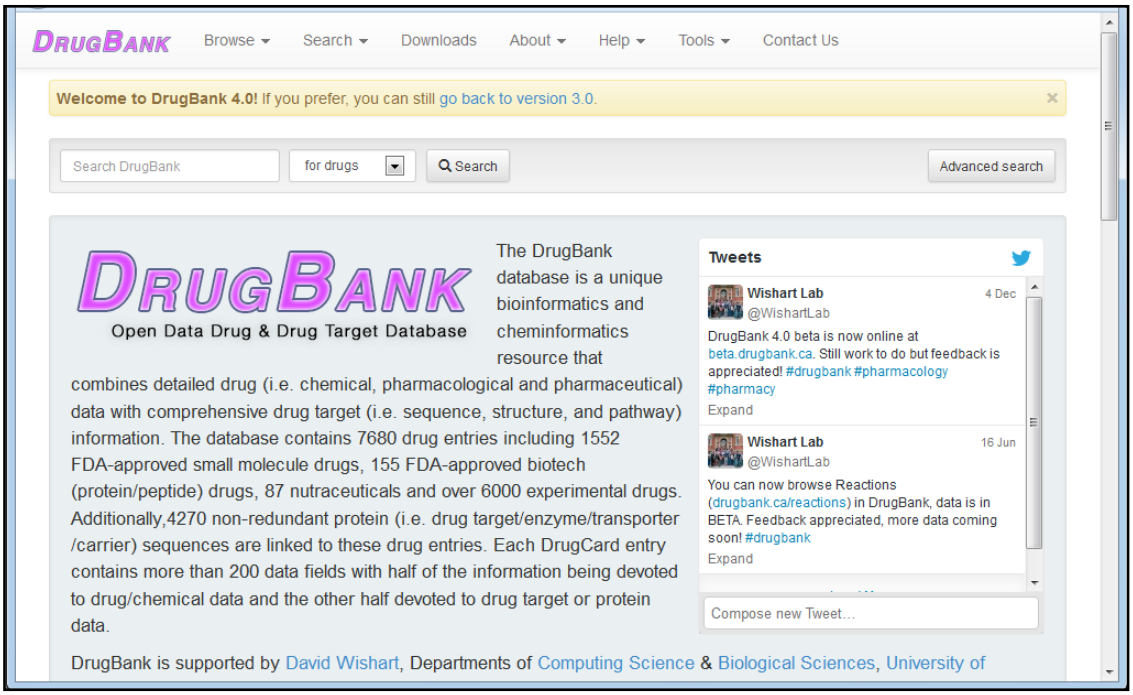
Zinc database: http://zinc.docking.org/

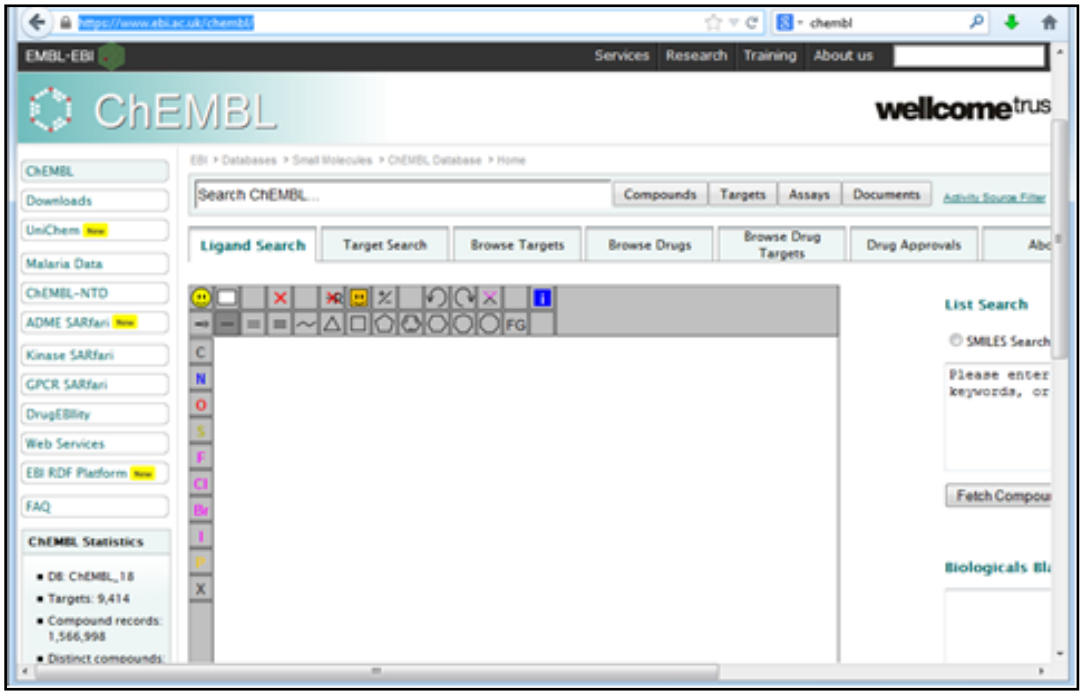
PubChem: https://pubchem.ncbi.nlm.nih.gov/

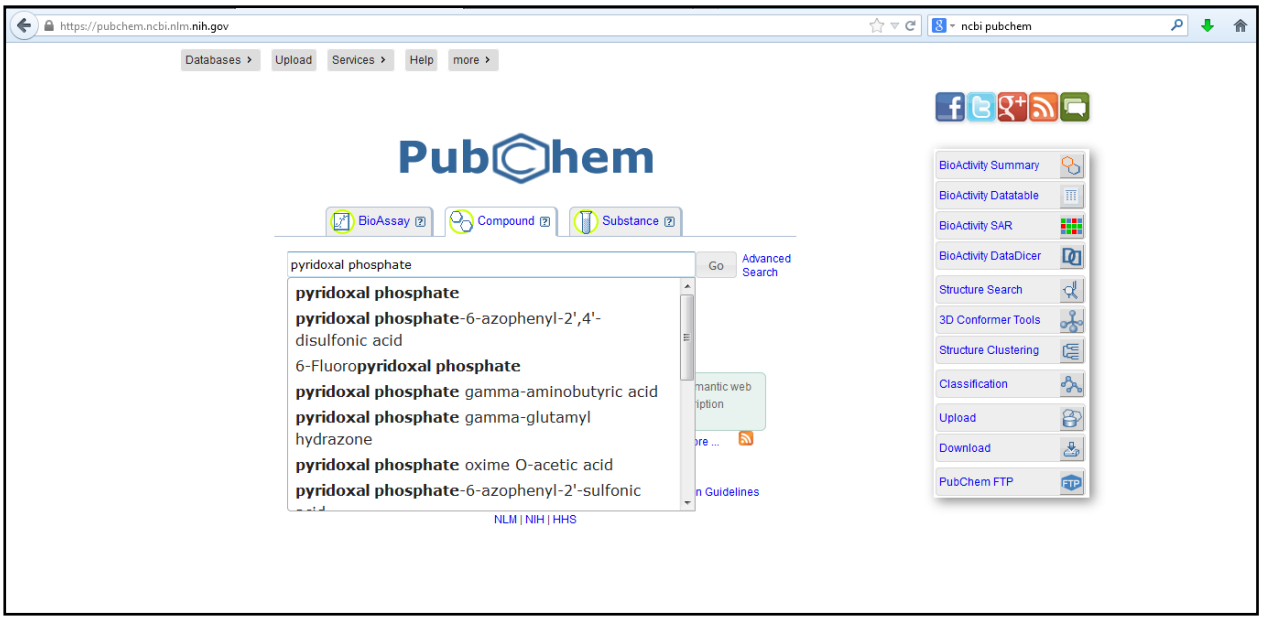
Drug bank: http://www.drugbank.ca/

ChemBL: https://www.ebi.ac.uk/chembl/




DrugBank
Browse
Search
Downloads
About
Help
Tools
Contact Us
Welcome to DrugBank 4.0! If you prefer, you can still go back to version 3.0.
Search DrugBank
for drugs
Search
Advanced search
DrugBank
Open Data Drug & Drug Target Database
The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. The database contains 7680 drug entries including 1552 FDA-approved small molecule drugs, 155 FDA-approved biotech (protein/peptide) drugs, 87 nutraceuticals and over 6000 experimental drugs. Additionally,4270 non-redundant protein (i.e. drug target/enzyme/transporter /carrier) sequences are linked to these drug entries. Each DrugCard entry contains more than 200 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data.
Tweets
Wishart Lab
@WishartLab
4 Dec
DrugBank 4.0 beta is now online at beta.drugbank.ca. Still work to do but feedback is appreciated! #drugbank #pharmacology #pharmacy
Expand
Wishart Lab
@WishartLab
16 Jun
You can now browse Reactions (drugbank.ca/reactions) in DrugBank, data is in BETA. Feedback appreciated, more data coming soon! #drugbank
Expand
Compose new Tweet…
DrugBank is supported by David Wishart, Departments of Computing Science & Biological Sciences, University of


EMBL-EBI
Services
Research
Training
About us
ChEMBL
wellcometrus
ChEMBL
Downloads
UniChem
Malaria Data
ChEMBL-NTD
ADME SARfari
Kinase SARfari
GPCR SARfari
DrugEBIlity
Web Services
EBI RDF Platform
FAQ
ChEMBL Statistics
Search ChEMBL...
Compounds
Targets
Assays
Documents
Ligand Search
Target Search
Browse Targets
Browse Drugs
Browse Drug Targets
Drug Approvals
List Search
SMILES Search
Fetch Compou
Biologicals Bl



https://pubchem.ncbi.nlm.nih.gov/

Montagem da biblioteca virtual

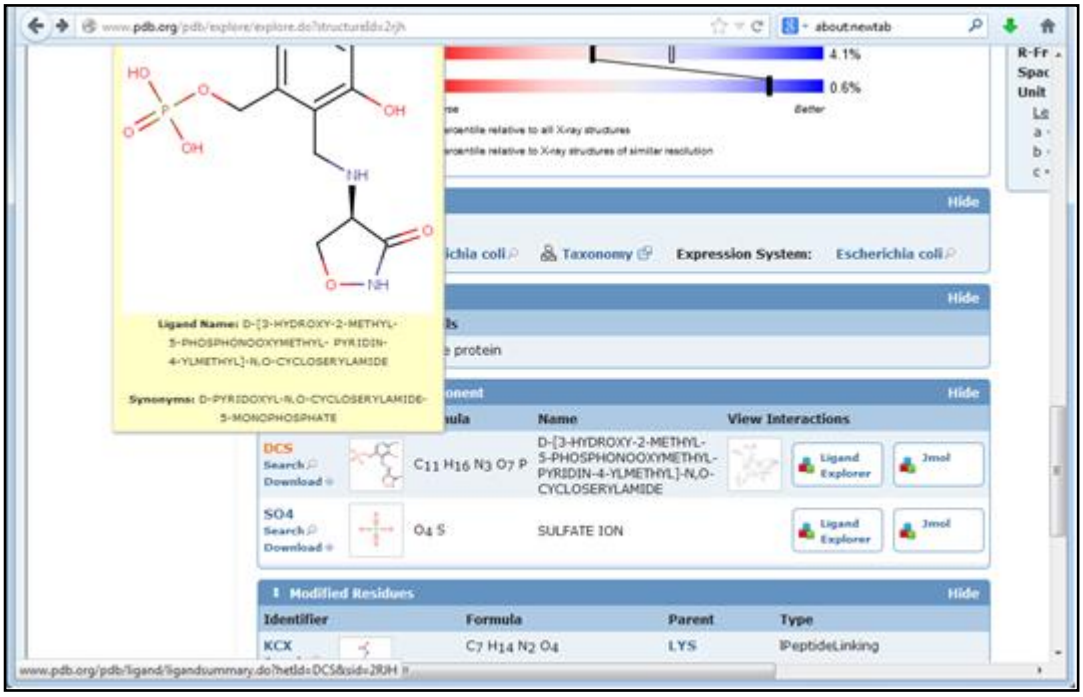
✓ **Objetivo:**

- Você deseja buscar um inibidor semelhante ao ligante utilizado na modelagem de sua enzima, o *D-Pyridoxyl-N,O-Cycloserylamide-5-Monophosphate*

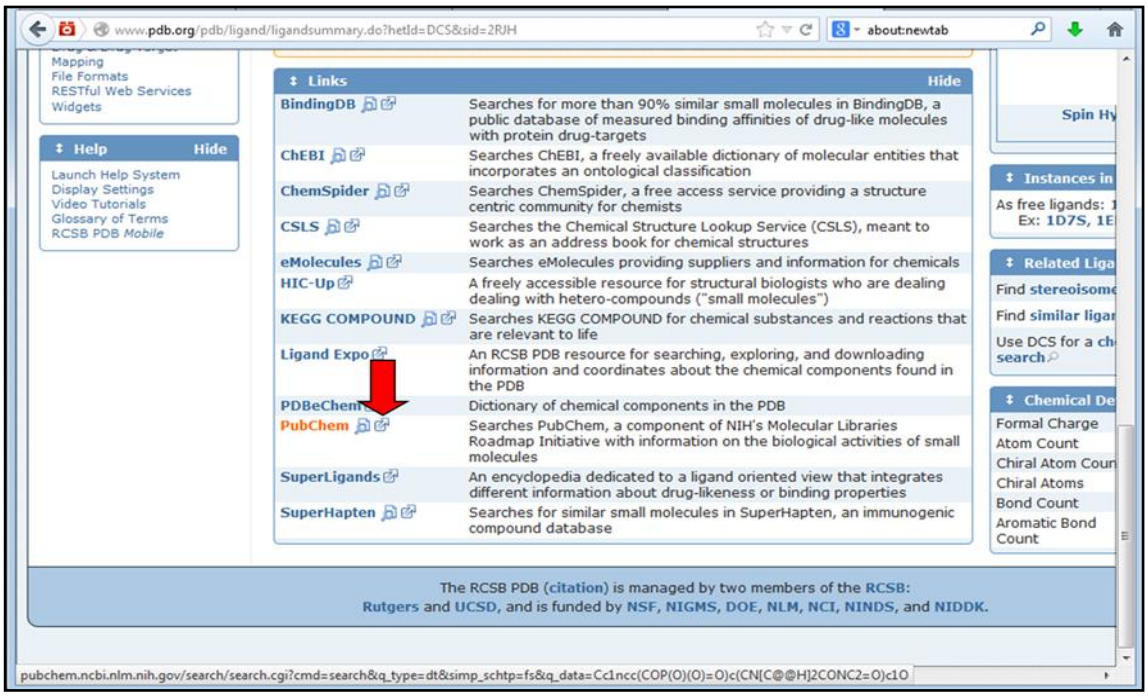
✓ **Metodologia:**

- Identificar o ligante na base de dados PubChem e depois buscar por compostos semelhantes a ele na base de dados Zinc.

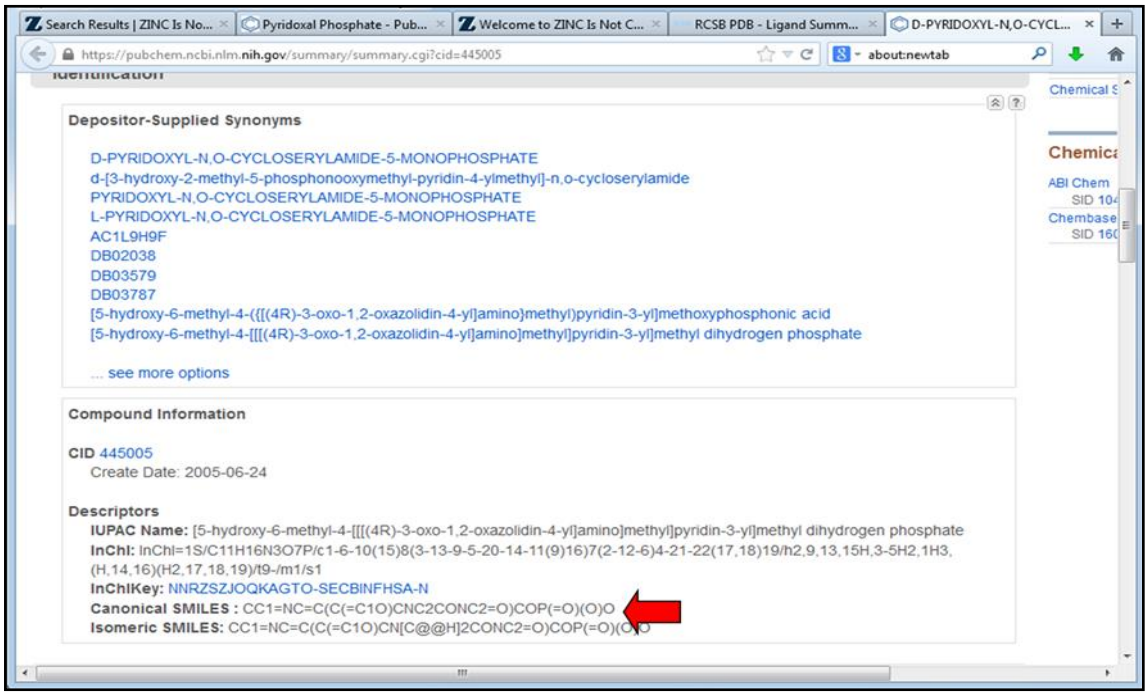
Ir no PDB (www.pdb.org), selecionar o molde 2RJH usado na modelagem. Depois ir até o ligante DCS e selecioná-lo.



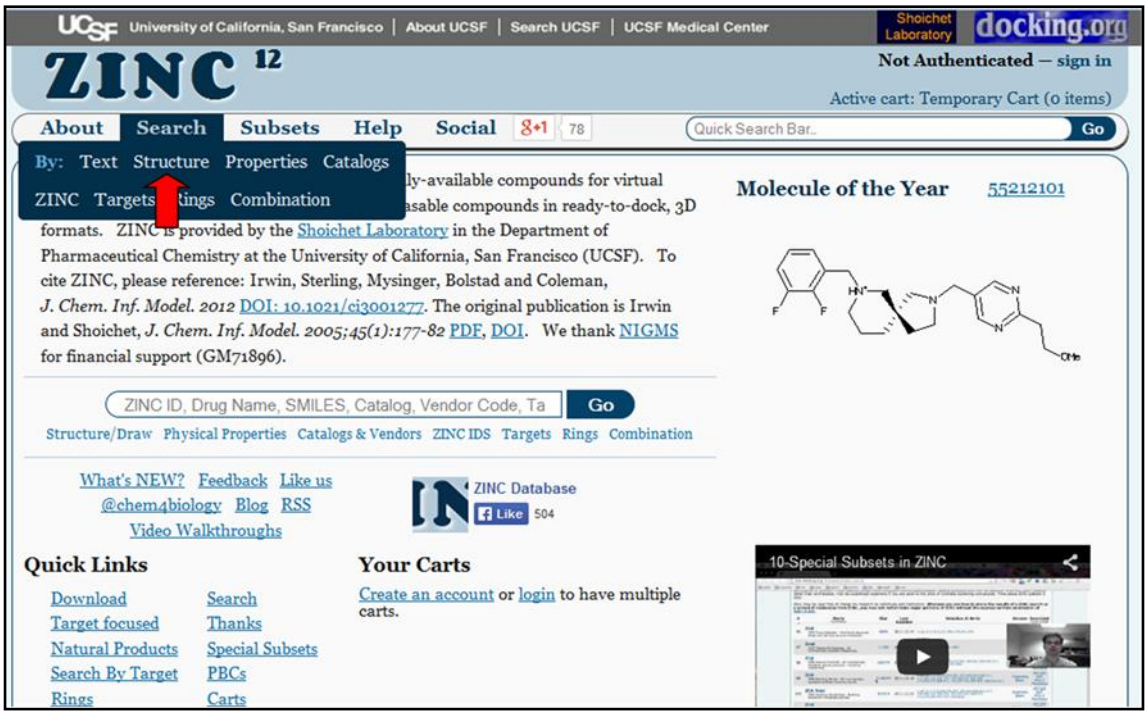
Na tela do ligante, selecionar o link para o Pubchem



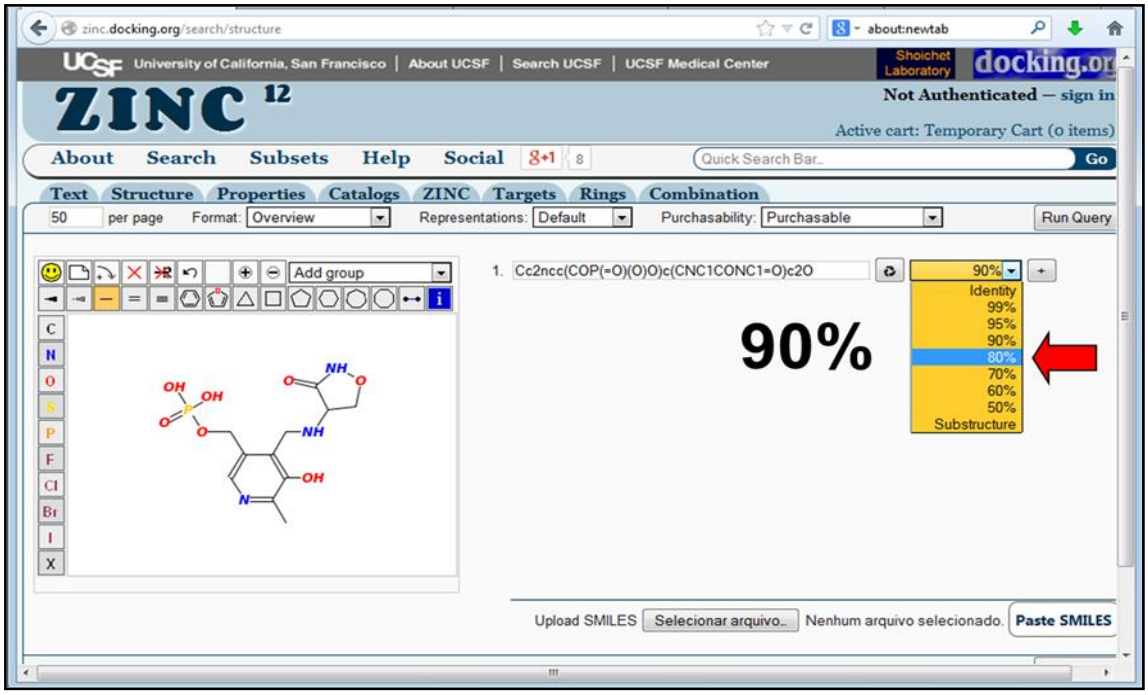
No PubChem, selecionar e copiar o código SMILES canônico



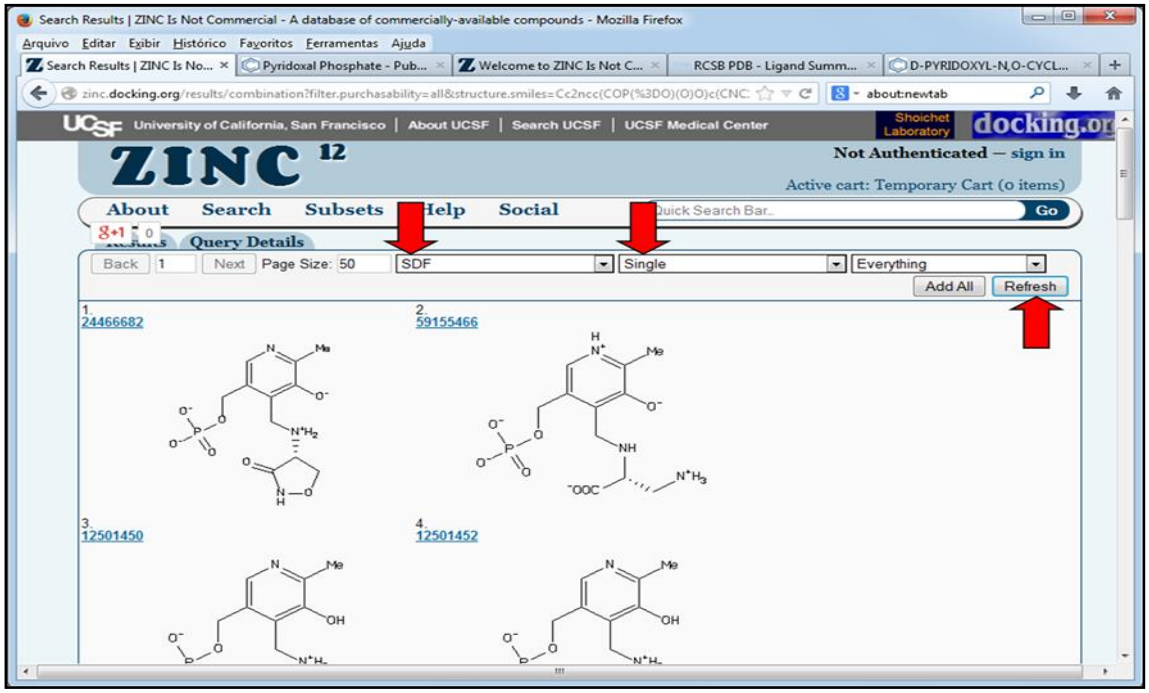
Em uma nova aba no navegador, vá para o Zinc database e selecionar Search → Structure



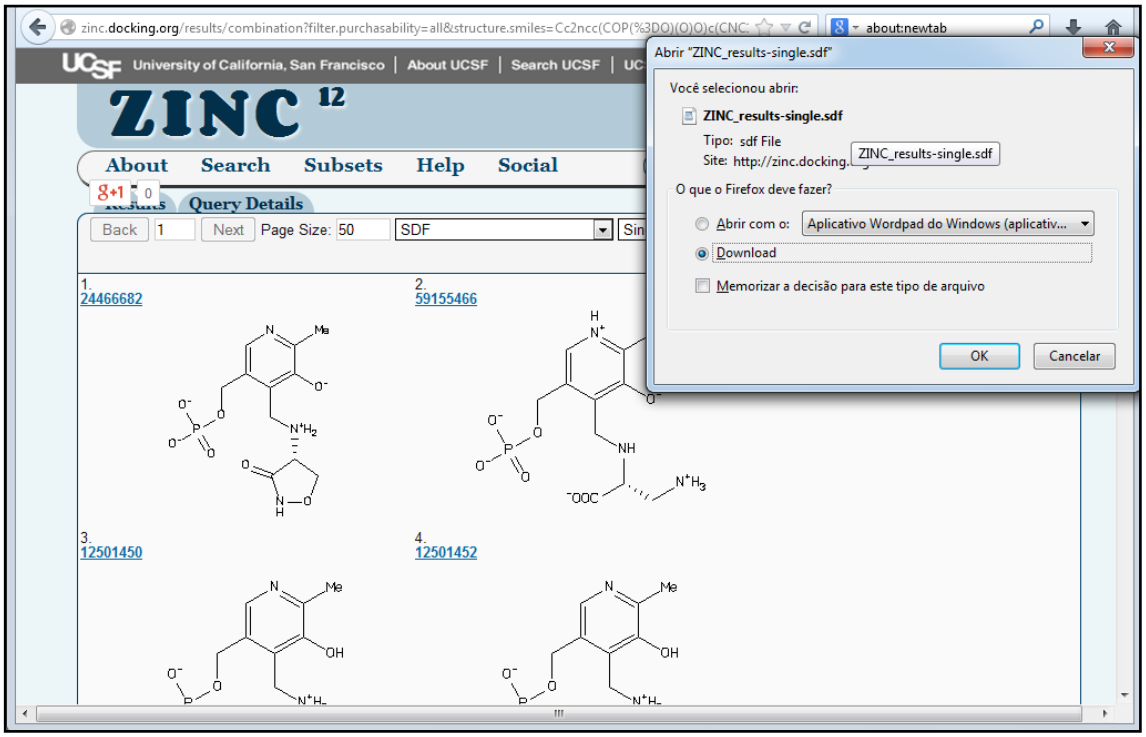
Colar o código SMILES e selecionar ligantes 90% similares



Download no formato SDF opção SINGLE e sem disponibilidade comercial (Everything)



Salve o resultado na pasta C:/Docking

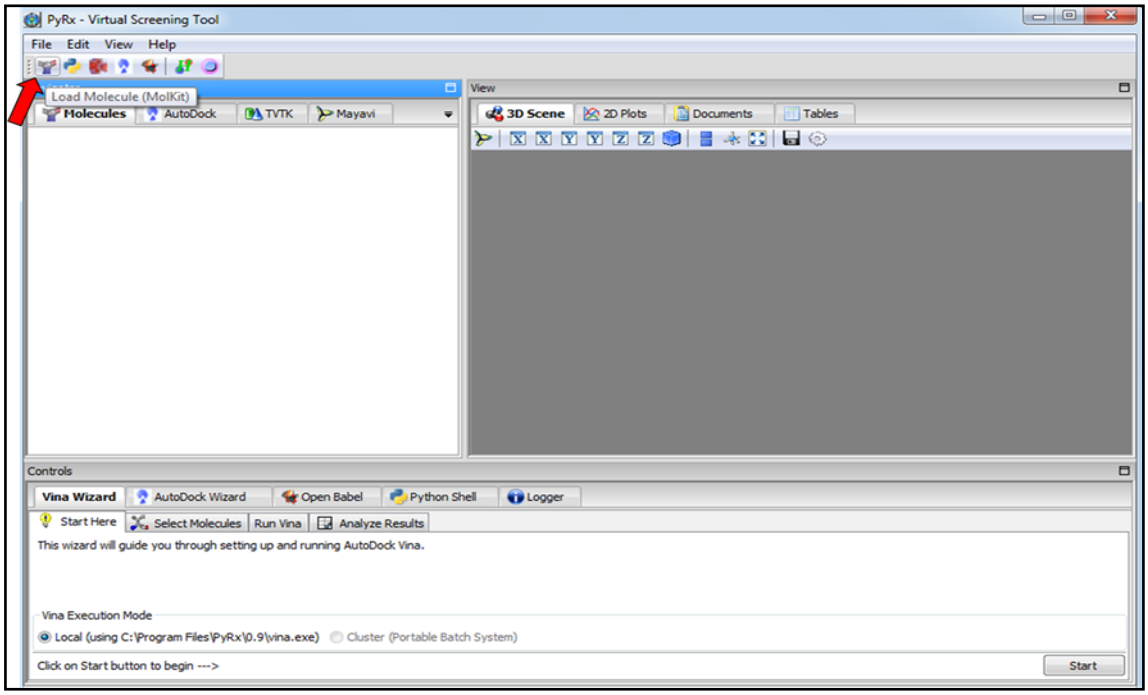


Varredura virtual (VS) da biblioteca montada.

O objetivo agora é usar o programa de docagem para varrer a biblioteca virtual que acabou de ser montada.

- ✓ Abra novamente o programa Pyrx,
- ✓ Carregue sua proteína modelada conforme feito anteriormente;
- ✓ Ajuste a cadeia A para Macromolecule
- ✓ Ajuste a cadeia B para Ligand

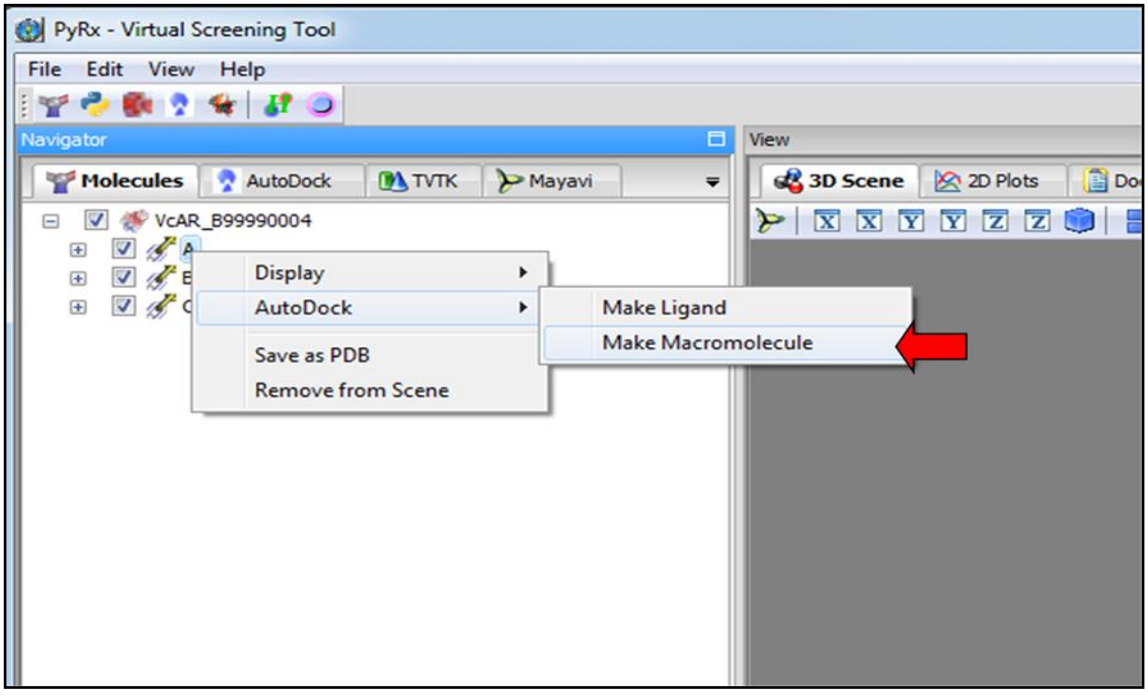
Carregar o programa, File → Load Molecule, depois, abra a estrutura modelada: VcAR.B99990004.pdb



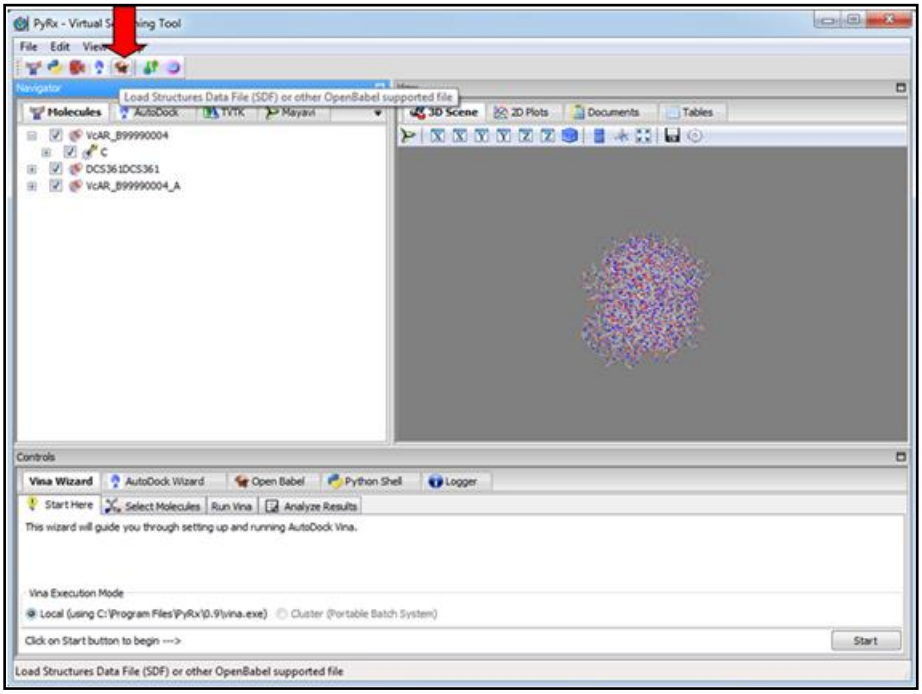
Clicar com o botão direito do mouse sobre as letras:

A → Autodock → Make Macromolecule

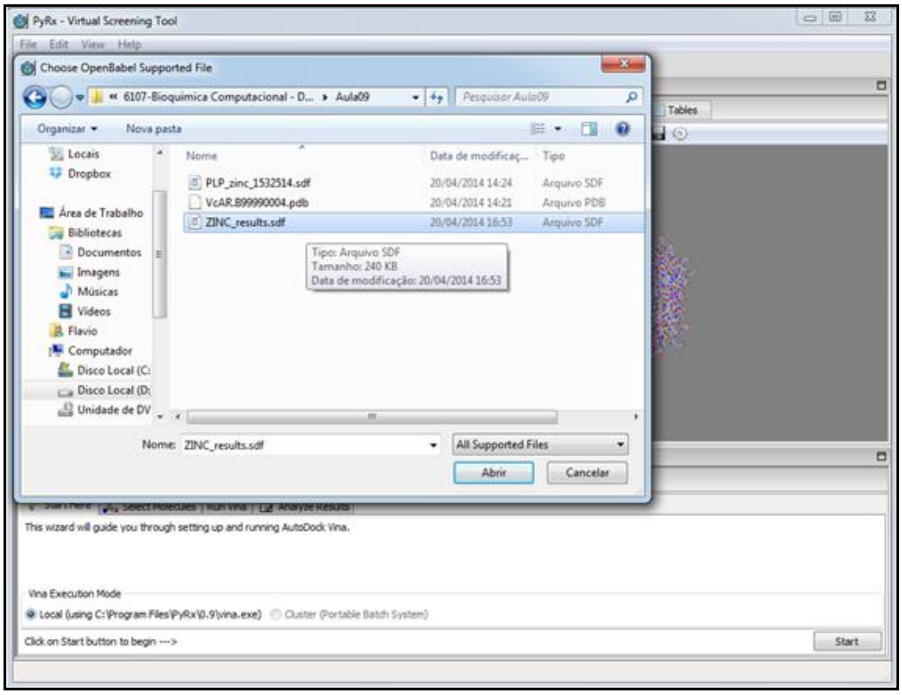
B → Autodock → Make Ligand



Depois de carregar a proteína e o ligante, carregar também a biblioteca

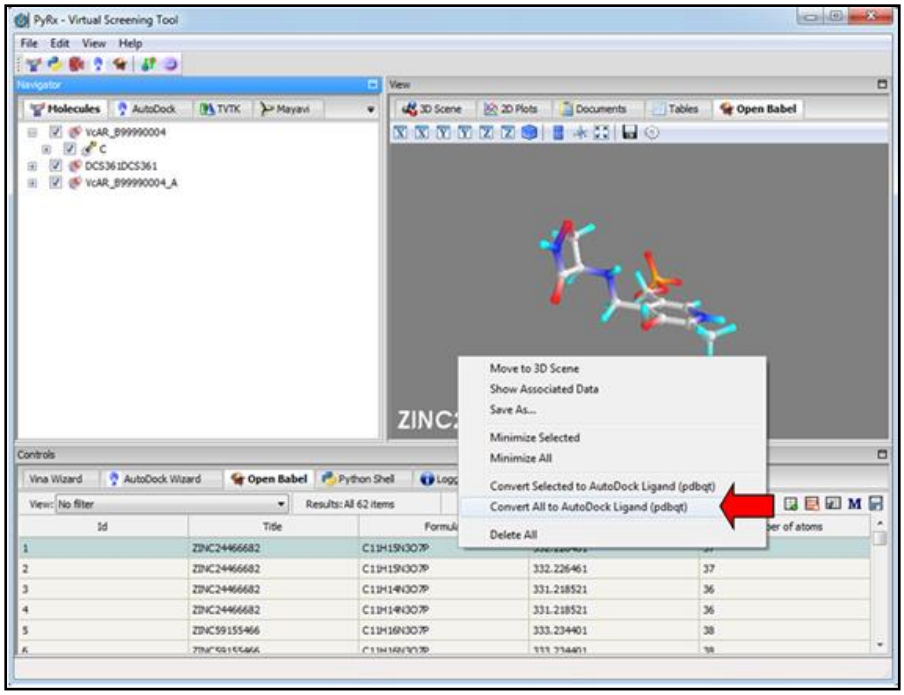


Procure na pasta C:/Docking o arquivo com as estruturas baixadas do Zinc database

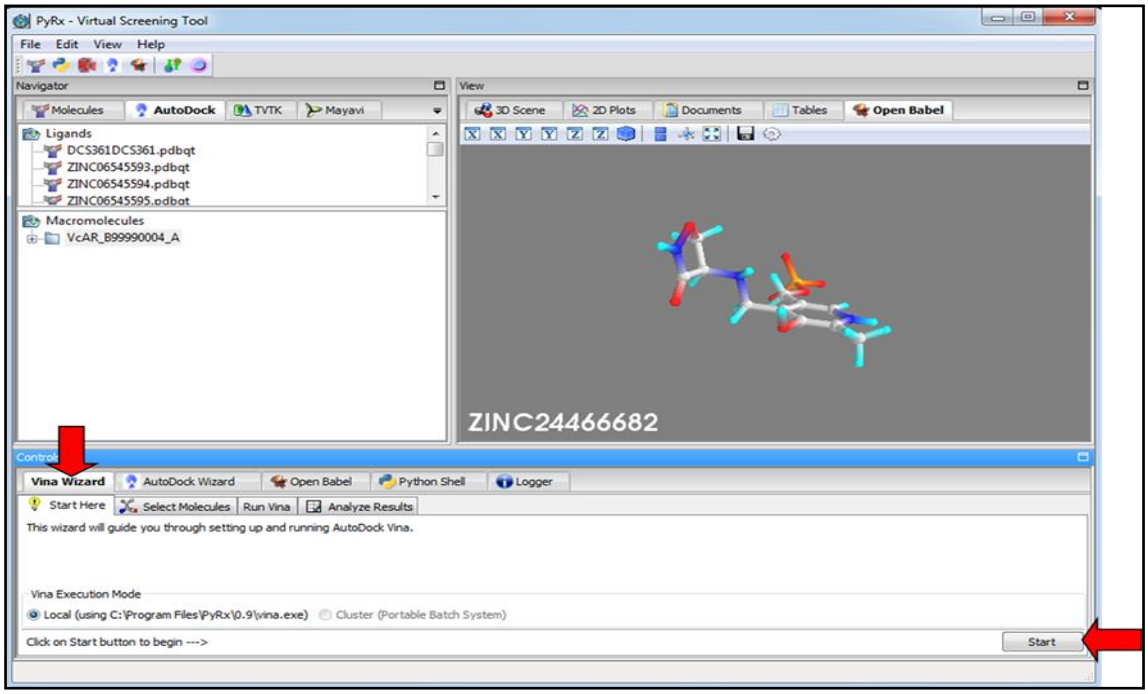


Clique com o botão direito sobre a fórmula molecular e escolha:

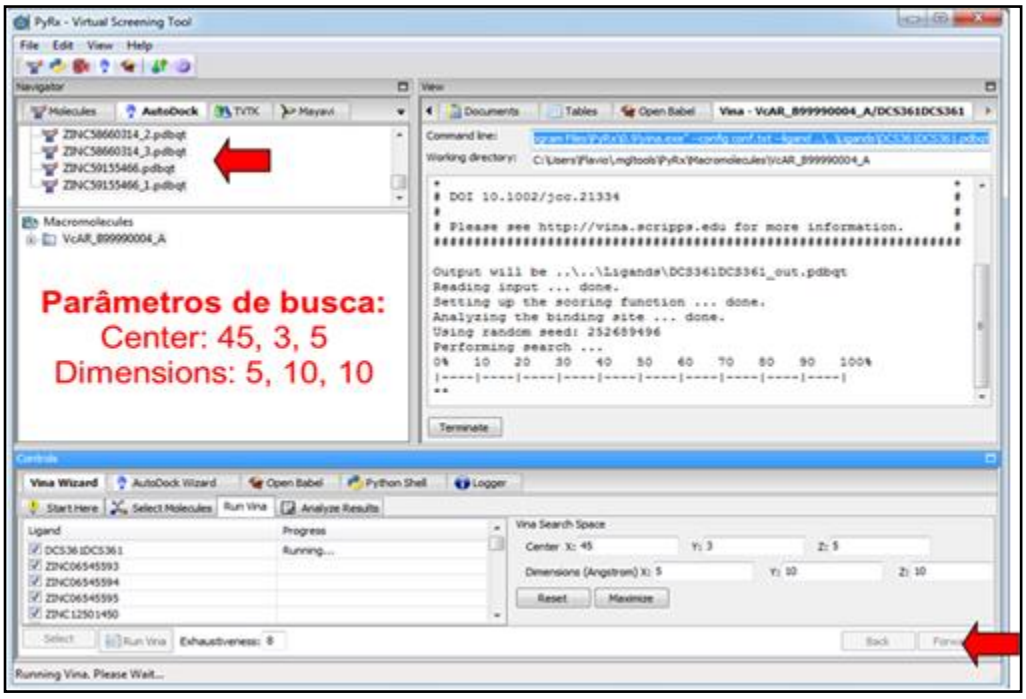
Convert all to Autodock Ligand (pdbqt)



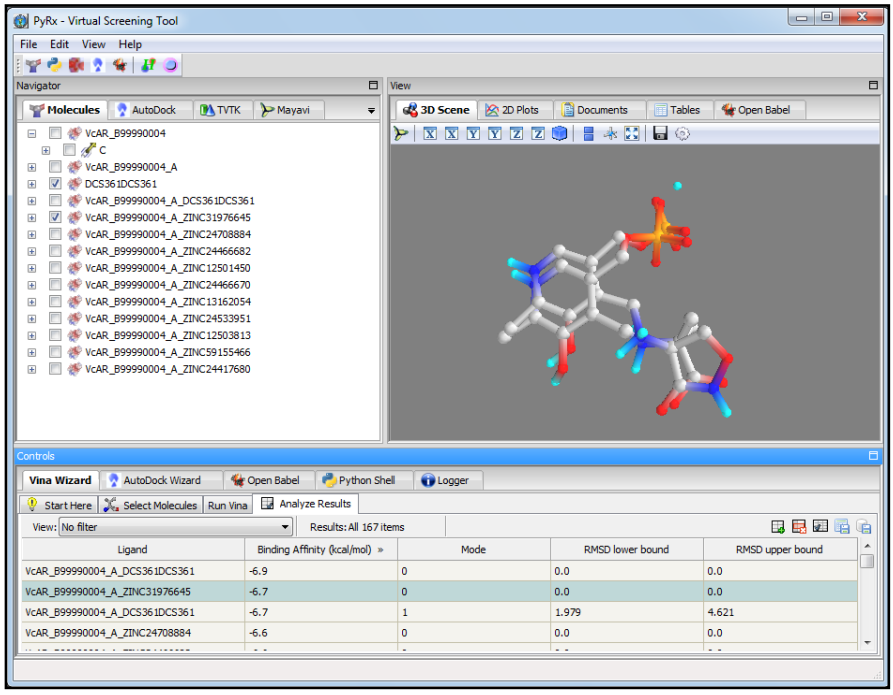
Depois de converter os ligantes, proceda conforme o protocolo validado no redocking.



Se necessário, selecione os ligantes e depois clique em *Forward*. Aguarde o resultado.



Verifique os ligantes que tiveram uma energia de interação $\Delta G_{binding}$ teórico, menor que o do ligante de referência.



O que fazer a partir daqui?

- Uma única simulação de docking não garante nada.
- É necessário repetir várias vezes a simulação para ver se há reprodutibilidade dos resultados, evitando-se assim, resultados falso-positivos.
- É aconselhável testar sua biblioteca em diferentes programas (que usam diferentes algorítmos) de docagem.
- A reprodutibilidade dos resultados com outros softwares fornece mais segurança para se iniciar a próxima fase da pesquisa.
- O próximo passo seria comprar (ou sintetizar) o ligante selecionado e testá-lo em cultura da bactéria *in vitro*.
- Em havendo atividade antimicrobiana (Minimal Inhibitory Concentration, MIC) em concentrações menores que 10 μM, é possível solicitar a patente dos produtos, e só aí publicar os resultados, pois há interesse da indústria farmacêutica em realizar novos testes *in vitro* e *in vivo* (pesquisa pré-clínica).

6 Dinâmica Molecular

6. 1 Introdução

Durante a última década, a simulação computacional aplicada a dinâmica de proteínas tornou-se uma ferramenta amplamente utilizada para aprofundar a compreensão no estudo destas moléculas. No método de simulação de dinâmica molecular (DM), equações de movimento de Newton para partículas ou átomos de um sistema molecular são resolvidas numericamente em função do tempo. Dada uma configuração inicial no sistema e as velocidades dos átomos, uma simulação de DM reproduz uma série de configurações espaciais possíveis em função do tempo, isto é, uma trajetória do sistema molecular. A partir de tal trajetória, as propriedades deste sistema podem ser calculadas (Gunsteren, 1993). As simulações de DM fornecem informações detalhadas sobre flutuações e mudanças conformacionais em proteínas e ácidos nucleicos. Estes métodos são rotineiramente utilizados para investigar a estrutura, dinâmica e termodinâmica de moléculas biológicas e seus complexos (Adcock & McCmmon, 2006).

O método de DM foi introduzido pela primeira vez por Alder e por Wainwright no final dos anos 50 (Alder & Wainwright, 1959) no estudo de interações entre esferas rígidas. Muitas informações importantes sobre o comportamento dos líquidos simples surgiram a partir de seus estudos. Outro grande avanço foi em 1964, quando Rahman realizou a primeira simulação usando um potencial realista para o argônio (AR) líquido (Rahman, 1964). A primeira simulação de um sistema realista foi realizada por Rahman e Stillinger na sua simulação de água líquida em 1974 (Rahman & Stillinger, 1974). As primeiras proteínas apareceram em simulações no ano de 1977, com a simulação do inibidor da tripsina pancreática bovina (McCammon, et al., 1977). Atualmente, na literatura encontram-se rotineiramente simulações de DM de proteínas solvatadas, complexos de proteína-DNA, bem como sistemas de lipídeos, abordando uma variedade de questões, incluindo a termodinâmica da ligação com o ligante, o dobramento de pequenas proteínas e a cinética enzimática. O número de técnicas de DM tem aumentado muito, hoje em dia existem técnicas especializadas para problemas específicos. As simulações clássicas são empregadas no estudo de reações enzimáticas no contexto da proteína completa, também são amplamente utilizadas técnicas de DM na análise de procedimentos experimentais, como a determinação de estruturas por cristalografia de raios-X e por RMN (Adcock & McCmmon, 2006).

As simulações podem demonstrar finos detalhes sobre os movimentos de partículas individuais em função do tempo. Elas podem ser utilizadas para quantificar as propriedades de

um sistema com precisão e em uma escala de tempo que de outro modo seria impossível, a simulação é, portanto, uma ferramenta valiosa no que diz respeito a nossa compreensão dos sistemas modelo. A abordagem teórica de um sistema permite, adicionalmente, a investigação das contribuições específicas de um átomo através da "alquimia computacional", isto é, modificando a simulação de uma forma não-física, mas ainda assim, permitindo as características de um modelo realista. Um exemplo em particular é a conversão artificial da função de energia de um sistema de representação a outro durante a simulação. Esta é uma técnica importante nos cálculos de energia livre. Assim, as simulações de DM, juntamente com diversas abordagens computacionais complementares, tornaram-se uma ferramenta fundamental na investigação básica da estrutura e função de proteínas (Adcock & McCmmon, 2006).

6. 1. 2 Aplicação da Dinâmica Molecular no Estudo de Fenômenos Biomoleculares

Atualmente a DM pode ser aplicada na investigação de diversas propriedades e processos dinâmicos por pesquisadores das mais diversas áreas, incluindo a bioquímica estrutural, a biofísica, a enzimologia, a biologia molecular, a química farmacêutica e a biotecnologia. Utilizando as simulações de DM pode-se estudar as propriedades termodinâmicas e os fenômenos dependentes do tempo (isto é, a cinética). Isso permite uma compreensão ampla de vários aspectos dinâmicos da estrutura biomolecular, como características de reconhecimento e sua função. No entanto, quando utilizada isoladamente, a DM é de utilidade limitada. Uma trajetória de DM (ou seja, o progresso da estrutura simulada em relação ao tempo), geralmente fornece dados apenas ao nível de posições atômicas, velocidades e energias de ponto único (Adcock & McCmmon, 2006). A metodologia da DM é fundamentada nos princípios da Mecânica Clássica e fornece informações sobre o comportamento dinâmico microscópico, dependente do tempo, dos átomos individuais que compõem o sistema. Para se obter as propriedades macroscópicas de interesse, a aplicação da mecânica estatística é requerida, a qual tem a função de calcular propriedades observáveis macroscópicas (pressão, energia interna, volume, temperatura, entropia, energia livre, etc), a partir de outras microscópicas (Namba et al., 2008).

Com base na Mecânica Molecular (MM), as moléculas são tratadas como uma coleção de átomos que pode ser descrita por forças newtonianas, ou seja, são tratadas como uma coleção de partículas mantidas unidas por forças harmônicas ou elásticas. Um conjunto completo dos potenciais de interação entre as partículas é referido como "campo de força". O campo de força

empírico, tal como é conhecido como uma função energia potencial, permite que a energia potencial total do sistema seja calculada como a partir da estrutura tridimensional (3D) deste sistema (Namba et al., 2008). Além disso, aspectos específicos da estrutura biomolecular, cinéticos e termodinâmicos, que podem ser investigados através da DM incluem, por exemplo, a estabilidade macromolecular, propriedades de conformação e sítios alostéricos, o papel da dinâmica na atividade enzimática, reconhecimento molecular e propriedades dos complexos, como íons e pequenas moléculas, associação a proteínas, dobramento de proteínas e a sua hidratação. A MD, portanto, possibilita uma grande diversidade de estudos, incluindo o design molecular (muito utilizado na concepção de fármacos), na determinação de estrutura e seu refinamento (raio-X, RMN e modelagem de proteínas) (Adcock & McCmmon, 2006).

6. 2 Etapas da simulação de DM

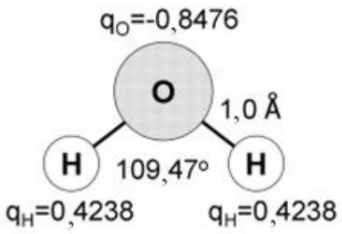
Os cálculos de sistemas com centenas ou até milhares de átomos passam por alguns passos para a completa modelagem do sistema, como descrito a seguir:

1) **Geração das configurações iniciais**: Podem ser obtidas em banco de dados especializados, ou geradas através de algum programa, quando não se dispõe da estrutura cristalográfica;
2) **Cálculo das forças exercidas sobre cada partícula**: Esta etapa é realizada em um *campo de força*;
3) **Otimização da estrutura**: Realizada através de algoritmos genéticos ou métodos de gradientes;
4) **Dinâmica da estrutura**: Realizada através da integração da equação de movimento de Newton, por métodos numéricos;
5) **Análise dos resultados** através das propriedades de equilíbrio.

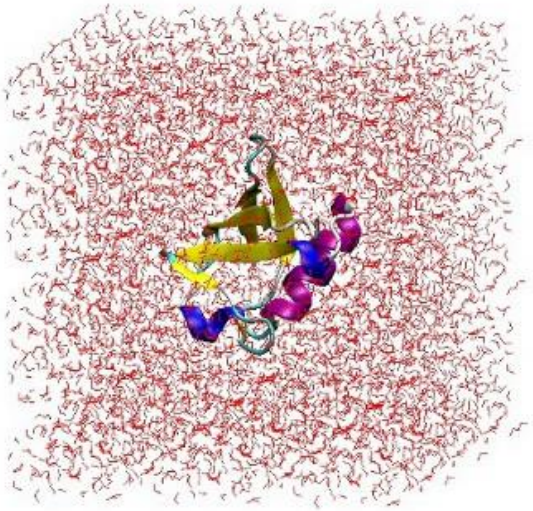
6. 2. 1 Configurações Gerais do Sistema

A determinação das coordenadas iniciais da molécula é o primeiro passo para a simulação do sistema. Podem ser utilizadas tanto estruturas cristalográficas baixadas do PDB, quanto modeladas (arquivo *.pdb). A partir das coordenadas espaciais é gerado um arquivo chamado *Protein Structure File* (*.psf), que contém todas as informações necessárias para aplicar um campo de força a um sistema molecular, exemplo: (quais átomos estão ligados, interagindo entre si, etc.).

Inserção de Solvente no Soluto: Uma das vantagens dos cálculos de dinâmica molecular é a possibilidade da inclusão do solvente de forma explícita, isto é, as moléculas estão realmente inseridas na simulação. Além do solvente este método permite a inclusão de íons para neutralizar cargas do sistema. O modelo mais usado nas simulações é do tipo *Simple Point Charge*. Este modelo descreve uma molécula de água com três sítios de interação, isto é, três pontos de carga (um no oxigênio, dois nos hidrogênios). Este modelo é considerado um dos modelos mais bem-sucedidos por sua simplicidade e eficiência computacional. Representar moléculas de água através de modelos pontuais dependem da parametrização do campo de força. Para aproximar o resultado experimental dos cálculos teóricos, este modelo considera o ângulo entre o oxigênio e os hidrogênios de 109, 47º. Este procedimento faz com que o momento de dipolo do modelo se aproxime do real.



Modelo de água tipo *Simple Point Charge* (SPC) usado nas simulações de DM.



Caixa d'água gerada pelo programa de DM.

Condições de Contorno Periódicas: Para a inclusão do efeito de solvatação no sistema a ser simulado, devem ser usadas condições de contorno periódicas, afim de que o sistema seja simulado como estivesse numa caixa de tamanho infinito.

6. 2. 2 Cálculo das Forças Exercidas Sobre Cada Partícula

Campos de Força

Os campos de força (*FF – Force fields*) existentes foram desenvolvidos de maneira independente e com todos os conjuntos de parâmetros específicos. Alguns incluem outros termos para descrever especificamente as ligações de hidrogênio ou para acoplar oscilações entre ângulos e comprimentos de ligação, com o objetivo de se obter uma melhor concordância com espectros vibracionais. A confiabilidade dos resultados é baseada na elaboração de um campo de força com parâmetros bem definidos. A escolha do campo de força depende, em grande parte, do sistema a ser estudado e das propriedades que serão investigadas. No caso de sistemas biomoleculares, os campos de força mais utilizados são CHARMM, GROMOS, AMBER, entre outros (Namba et al., 2008).

O campo de força é dividido em contribuições intermoleculares e intramoleculares cuja soma descreve a energia potencial total do sistema. As penalizações de energia potencial sofridas pelo sistema estão associadas ao desvio dos valores de referência obtidos experimentalmente ou por métodos de mecânica quântica.

Os campos de força são subdivididos em dois arquivos: Um com a topologia (top_all22_prot.top) e outro contendo os parâmetros (par_all22_prot.par).

(CHARMM *force fields*: http://mackerell.umaryland.edu/CHARMM_ff_params.html)

Arquivo de topologia: **top_all22_prot.top**

- ✓ Diz respeito a nomenclatura, massa e cargas parciais dos átomos e dos resíduos de aminoácidos:

```
MASS     1 H      1.00800 H ! polar H
MASS     2 HC     1.00800 H ! N-ter H
MASS     3 HA     1.00800 H ! nonpolar H
MASS    20 C     12.01100 C ! carbonyl C, peptide backbone
MASS    21 CA    12.01100 C ! aromatic C
MASS    22 CT1   12.01100 C ! aliphatic sp3 C for CH
MASS    50 N     14.00700 N ! proline N
MASS    51 NR1   14.00700 N ! neutral his protonated ring nitrogen
MASS    52 NR2   14.00700 N ! neutral his unprotonated ring nitrogen
MASS    70 O     15.99900 O ! carbonyl oxygen
MASS    71 OB    15.99900 O ! carbonyl oxygen in acetic acid
MASS    72 OC    15.99900 O ! carboxylate oxygen
MASS    73 OH1   15.99900 O ! hydroxyl oxygen
MASS    81 S     32.06000 S ! sulphur
MASS    82 SM    32.06000 S ! sulfur C-S-S-C type
MASS    83 SS    32.06000 S ! thiolate sulfur
```

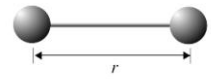
```
RESI ARG          1.00
GROUP
ATOM N    NH1    -0.47  !     |                       HH11
ATOM HN   H       0.31  !  HN-N                        |
ATOM CA   CT1     0.07  !     |   HB1 HG1 HD1 HE     NH1-HH12
ATOM HA   HB      0.09  !     |   |   |   |   |    //(+)
GROUP                   !  HA-CA--CB--CG--CD--NE--CZ
ATOM CB   CT2    -0.18  !     |   |   |   |          \
ATOM HB1  HA      0.09  !     |   HB2 HG2 HD2        NH2-HH22
ATOM HB2  HA      0.09  !   O=C                        |
GROUP                   !     |                       HH21
ATOM CG   CT2    -0.18
ATOM HG1  HA      0.09
ATOM HG2  HA      0.09
```

Exemplos de informações contidas em um arquivo de topologia do campo de força CHARMM.

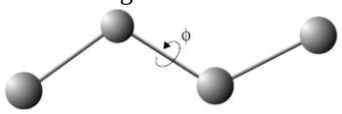
Arquivo de parâmetros: **par_all22_prot.par**

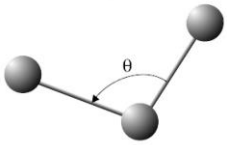
- ✓ Diz respeito às interações ligadas (distâncias e ângulos) e não ligadas (pontes de H, contatos de VdW) formadas entre os átomos:

Distância

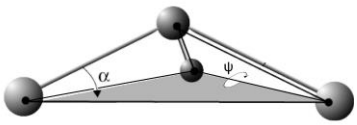


Ângulo diédrico





Ângulo



Ângulo diédrico "impróprio" ψ

```
BONDS
!
!V(bond) = Kb(b - b0)**2
!
!Kb: kcal/mole/A**2
!b0: A
!
!atom type Kb          b0
!
!Carbon Dioxide
CST  OST   937.96      1.1600 ! JES
!Heme to Sulfate (PSUL) link
SS   FE    250.0       2.3200 !force constant a guess
           !equilbrium bond length optimized to reproduce
           !CSD survey values of
           !2.341pm0.01 (mean, standard error)
           !adm jr., 7/01
C    C     600.000     1.3350 ! ALLOW ARO HEM
                ! Heme vinyl substituent (KK, from propene (JCS))
CA   CA    305.000     1.3750 ! ALLOW   ARO
                ! benzene, JES 8/25/89
CE1  CE1   440.000     1.3400   !
! for butene; from propene, yin/adm jr., 12/95
CE1  CE2   500.000     1.3420   !
```

```
ANGLES
!
!V(angle) = Ktheta(Theta - Theta0)**2
!
!V(Urey-Bradley) = Kub(S - S0)**2
!
!Ktheta: kcal/mole/rad**2
!Theta0: degrees
!Kub: kcal/mole/A**2 (Urey-Bradley)
!S0: A
!
!atom types     Ktheta    Theta0   Kub     S0
!
!Carbon Dioxide, JES
OST  CST  OST    3000.00  180.0000 ! CO2, JES
!Heme to Sulfate (PSUL) link
CS   SS   FE     50.0      100.6    !force constant a guess
                 !equilibrium angle optimized to reproduce
                 !CSD survey values
                 !107.5pm0.6 (mean, standard error)
                 !adm jr., 7/01
SS   FE   NPH    100.0      90.0    !force constant a guess
                 !adm jr., 7/01
!
CA   CA   CA     40.000    120.00   35.00   2.41620 ! ALLOW   ARO
```

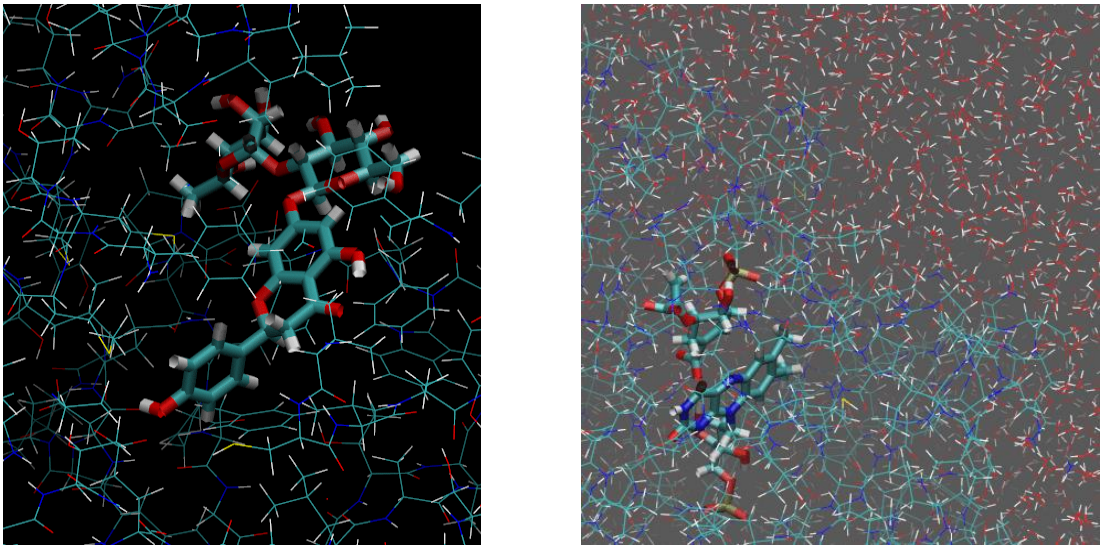
```
DIHEDRALS
!
!V(dihedral) = Kchi(1 + cos(n(chi) - delta))
!
!Kchi: kcal/mole
!n: multiplicity
!delta: degrees
!
!atom types             Kchi    n   delta
!
!Heme to Sulfate (PSUL) link
X    FE   SS   X        0.0000  4     0.00 ! guess
                !adm jr., 7/01
X    CS   SS   X        0.0000  3     0.20 ! guess
                !from methanethiol, HS S CT3 HA
                !adm jr., 7/01

C    CT1  NH1  C        0.2000  1   180.00 ! ALLOW PEP
                ! ala dipeptide update for new C VDW Rmin, adm jr., 3/3/93c
C    CT2  NH1  C        0.2000  1   180.00 ! ALLOW PEP
                ! ala dipeptide update for new C VDW Rmin, adm jr., 3/3/93c
C    N    CP1  C        0.8000  3     0.00 ! ALLOW PRO PEP
                ! 6-31g* AcProNH2, ProNH2, 6-31g*//3-21g AcProNHCH3 RLD 4/23/93
CA   CA   CA   CA       3.1000  2   180.00 ! ALLOW   ARO
                ! JES 8/25/89
CA   CPT  CPT  CA       3.1000  2   180.00 ! ALLOW   ARO
                ! JWK 05/14/91 fit to indole
```

Exemplos de informações contidas em um arquivo de parâmetros do campo de força CHARMM.

6. 2. 3 Otimização da Estrutura

Otimização da geometria: Método também conhecido como minimização da estrutura. Para otimizar os comprimentos e os ângulos das ligações, bem como as interações não ligadas utilizam-se métodos de gradiente. O mais utilizado é o *steepest descent* ou método de Caunchy. Sua desvantagem é que o resultado fica "preso" em mínimos locais de energia. Outro método empregado na otimização da geometria é o chamado gradiente conjugado. A grande vantagem do método de gradiente conjugado é que a direção do gradiente no próximo passo é sempre ortogonal ao ponto anterior. Este fato leva a direção de encontro ao mínimo sempre ortogonal à força aplicada.



Método de minimização por gradiente conjugado utilizando o programa NAMD2.

6. 2. 4 Dinâmica da Estrutura

Após a otimização da geometria molecular podemos estudar a evolução temporal do sistema em questão. Esta metodologia denominada dinâmica molecular, permite a simulação de um determinado sistema, em uma temperatura e pressão de interesse. Através da dinâmica molecular podemos gerar sucessivas configurações do sistema, integrando as equações do movimento de Newton. O resultado são trajetórias que especificam as variações das posições e velocidades com o tempo.

Com as novas posições e velocidades de cada partícula, obtêm-se as energias potencial e cinética do sistema. Aplicando-se sucessivamente esse procedimento, obtém-se o que se denomina de "trajetória", que nada mais é do que o conjunto de posições e velocidades de cada

partícula ao longo do tempo. Durante a simulação também são utilizados diferentes algoritmos para o controle da temperatura e da pressão.

6. 2. 5 Análise dos Resultados

No decorrer da simulação por dinâmica molecular, são realizadas duas fases do processo. A primeira denomina-se fase de equilíbrio e a segunda fase de produção. Na fase de equilíbrio os átomos pesados da estrutura são restritos por um potencial harmônico de constante 1000KJ(mol/nm^2) enquanto o solvente é relaxado em torno da estrutura. Após esta fase inicial todo o sistema pode movimentar-se livremente até atingir uma conformação de equilíbrio, onde as propriedades termodinâmicas de interesse são medidas. Algumas funções podem esclarecer muitas propriedades estruturais, tais como, distribuição atômica, diferença de conformação compactação de uma determinada estrutura.

- ✓ **Raiz quadrada média da distância (*RMSD*):** Em estudos de nanoestruturas é de grande interesse avaliar a diferença estrutural entre duas moléculas. No caso de fármacos estas diferenças estruturais podem corresponder a um polimorfismo que poderia levar um medicamento a ser tóxico ou a não ter a ação desejada. Pode-se também fazer uma média de RMSD durante a trajetória de simulação, avaliando assim a oscilação do sistema durante o tempo decorrido de uma determinada dinâmica. Esta média é de grande importância, porque pequenas oscilações de RMSD correspondem a posições de equilíbrio do sistema, enquanto mudanças bruscas na média denotam mudanças importantes na conformação de determinada molécula.

- ✓ **Raio de Giro (*RGyr*):** É a raiz quadrada média da distância entre o centro de gravidade da proteína com determinado átomo.

Referências:

Adcock, S. A.; McCammon, A. J. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. **Chem. Rev.**, v.106, n.5, p.1589-1615, 2006.

Alberts, B. *et al*. **Biologia Molecular da Célula**. 4a ed. Artmed editora, Porto Alegre, 2004.

Alder, B. J.; Wainwright, T. E. Studies in Molecular Dynamics. I. General Method. **J. Chem. Phys**., v.31, n.459, 1959.

Babu, M. M. **Biological databases and protein sequence analysis**. Center of Biotechnology Anna University, 1997.

Chen, Y. P. **Bioinformatics Technologies**. Berlin: Springer, 396p., 2005.

Cravedi, K. GenBank Celebrates 25 Years of Service with Two-Day Conference; Leading Scientists Will Discuss the DNA Database at April 7-8 Meeting. National Center for Biotechnology Information – **NCBI**, 2008.

David L. N., Cox M. M. **Princípios de Bioquímica de Lehninger**. 5. Ed. Porto Alegre: Artmed, p. 530-545, 2011.

Docagem Molecular – Wikipédia. Disponível em: <http://pt.wikipedia.org/wiki/Docagem_Molecular>

Edman, P.; Begg, G. A protein sequenator. **Eur. J. Biochem**., v.1, p.80-91, 1967.

Fenstermacher, David. Introduction to Bioinformatics. **J. Am. Soc. Inf. Sci. Tec.**, v.56, n.5, p. 440-446, 2005.

Fleischmann, R. D.; Adams, O.; White, R. A.; Clayton, E. F.; Kirkness, A. R.; Kerlavage, C.; Bult, J.; Tomb, J. F.; Dougherty, B. A.; Merrick, J. M.; *et al*. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. **Science**, v.28, p. 496-512, 1995.

Franco, María Liliana; Cediel, Juan Fernando; Payán, César. Breve historia de la bioinformática. **Colomb. Med.**, n.39, p.117-120, 2008.

Fruton, J. S. A. **Skeptical Biochemist**. Cambridge: Harvard Univ. Press., 1992.

Gunsteren, W. F. V. Molecular Dynamics Studies of Proteins. **Curr. op. in Struct. Biol.**, v.3, p.277-281, 1993.

Hagen, Joel. B. The origins of bioinformatics. **Nature Reviews: Genetics**, v.1, p.231-236, 2000.

Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G. Wyckoff, H. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. **Nature**, v.181, p.662-666, 1958.

Kendrew, J. C.; Dickerson, R. E.; Strandberg, B. E.; Hart, R. G.; Davies, D. R.; Phillips, D. C.; Shore, V. C. Structure of Myoglobin: A Three-Dimensional Fourier Synthesis at 2 Å. Resolution. **Nature**, v.185, p.422-427, 1960.

Lengauer T, Rarey M. Computational methods for biomolecular docking. **Current Opinion in Structural Biology**, v 6, n 3, p 402–406, 1996.

Lesk, A. M. **Introduction to Protein Architecture.** Oxford University Press, New York, 2001.

Mathews, C. K.; Van Holde, K. E.; Appling, D. R.; Anthony-Cahill, S. J. **Biochemistry**, 4 ed., Prentice Hall, 2012.

Mathura, V. S.; Kangueane, P. Bioinformatics: **A concep-based introduction**. New York: Springer, 184p., 2009.

McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of folded proteins. **Nature**, v.267, n.5612, p.585-590, 1977.

Modeller software (download gratuito mediante licença) <http://salilab.org/modeller/>

Namba, A. M.; Silva, V. B.; Silva, C. H. T. P. Dinâmica Molecular: Teoria e aplicações em Planejamento de Fármacos. **Ecl. Quím**., v.33, n.4, p.13-24, 2008.

Perutz, M. F.; Rossmann, M. G.; Cullis, A. F.; Muirhead, H. Will, G.; North, A. C. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å. resolution, obtained by X-ray analysis. **Nature**, v.185, p.416-422, 1960.

Prosdocimi, F. **Curso On Line: Introdução à Bioinformática**. [S.I.]: Portal Biotecnologia, 2007. Disponível em: <http://www2.bioqmed.ufrj.br/prosdocimi/FProsdocimi07_CursoBioinfo.pdf>. Acesso em: 26 jun. 2016.

Pyrx Screencast (vídeos tutoriais em Inglês). Disponível em: <http://pyrx.sourceforge.net/videos>

Rahman, A. Correlations in the Motion of Atoms in Liquid Argon. **Phys. Rev**., v.136, n.2, p.405-411, 1964.

Raza, K. Formal Concept Analysis for Knowledge Discovery from Biological Data. **ArXiv e-prints**, 2015. Disponível em: < http://arxiv.org/pdf/1506.00366v1.pdf>. Acesso em: 27 jun. 2016.

Ryle, A. P.; Sanger, F.; Smith, L. F.; Kitai, R. Insulin amide groups. **Biochem J**., v.60, p541-556, 1955.

Sanger, F.; Air, G. M.; Barrell, B. G.; Brown, N. L.; Coulson, A. R.; Fiddes, J. C.; Hutchison, C. A.; Slocombe, P. M.; Smith, M. Nucleotide sequence of bacteriophage φX174 DNA. **Nature**, v.265, p.687-695, 1977.

Stillinger, F. H.; Rahman, A. Improved simulation of liquid water by molecular dynamics. **J. Chem. Phys.**, v.60, n.1545, 1974.

Tateno, Y.; Manishi, T.; Miyazaki, S.; Fukami-Kobayashi, K.; Saitou, N.; Sugawara, H.; Gojobori, T. The DNA Data Bank of Japan (DDBJ) for genome scale research in life sciences. **Nucleic Acid Res**., v.30, n.1, p. 27-30, 2002.

Thampi, Sabu M. Introduction to bioinformatics. **ArXiv e-prints**, 2009. Disponível em: <https://arxiv.org/ftp/arxiv/papers/0911/0911.4230.pdf>. Acesso em: 28 jun. 2016.

Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; Gocayane, J. D.; Amanatides, P.; Ballew, R. M.; Huson, D. H.; Wortman, J. R.; et al. The Sequence of The Human Genome. **Science**, v.291, n.5507, p.1304-1351, 2001.

Verli, H. **Bioinformática: Da biologia à flexibilidade molecular**. Porto Alegre: *e-book*. Disponível em: <http://www.ufrgs.br/bioinfo/ebook/>. Acesso em: 28 jun. 2016.

Weeler, D. L.; Church, D. M.; Lash, A. E.; Leipe, D. D.; Madeen, T. L.; Pontius, J. U.; Schuler, G. D.; Schriml, L. M.; Tatusova, T. A.; Wagner, L.; Rapp, B. A. Database resources of the National Center for Biotechnology information: 2002 update. **Nucleic Acid Res**., v.30, n.1, p.13-16, 2002.