ÁLLAN JHONATHAN RAMOS FERRARI

CARACTERIZAÇÃO ESTRUTURAL DA STANNIOCALCINA-1 POR PROTEÔMICA ESTRUTURAL

CAMPINAS

2015

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE QUÍMICA

ÁLLAN JHONATHAN RAMOS FERRARI

CARACTERIZAÇÃO ESTRUTURAL DA STANNIOCALCINA-1 POR PROTEÔMICA ESTRUTURAL

ORIENTADOR: PROF. DR. FÁBIO CESAR GOZZO

DISSERTAÇÃO DE MESTRADO APRESENTADA AO INSTITUTO DE QUÍMICA DA UNICAMP PARA OBTENÇÃO DO TÍTULO DE MESTRE EM QUÍMICA NA ÁREA DE QUÍMICA ORGÂNICA.

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA POR ÁLLAN JHONATHAN RAMOS FERRARI, E ORIENTADA PELO PROF. DR. FABIO CESAR GOZZO.

_____

Assinatura do Orientador

CAMPINAS

2015

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Química
Simone Lucas Gonçalves de Oliveira - CRB 8/8144

F412c

Ferrari, Állan Jhonathan Ramos, 1991-
Caracterização Estrutural da Stanniocalcina-1 por Proteômica Estrutural / Állan Jhonathan Ramos Ferrari. – Campinas, SP : [s.n.], 2015.

Orientador: Fábio Cesar Gozzo.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Química.

1. Espectrometria de massas. 2. Proteômica estrutural. 3. Ligação cruzada. 4. Proteínas. I. Gozzo, Fabio Cesar. II. Universidade Estadual de Campinas. Instituto de Química. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Structural characterization of Stanniocalcin-1 by Structural Proteomics
Palavras-chave em inglês:
Mass spectrometry
Structural proteomics
Crosslinking
Proteins
Área de concentração: Química Orgânica
Titulação: Mestre em Química na área de Química Orgânica
Banca examinadora:
Fábio Cesar Gozzo [Orientador]
Carlos Henrique Inacio Ramos
Vitor Marcelo Silveira Bueno Brandão de Oliveira
Data de defesa: 24-02-2015
Programa de Pós-Graduação: Química

"Somos assim. Sonhamos o voo, mas tememos as alturas. Para voar é preciso amar o vazio. Porque o voo só acontece se houver o vazio. O vazio é o espaço da liberdade, a ausência de certezas. Os homens querem voar, mas temem o vazio. Não podem viver sem certezas. Por isso trocam o voo por gaiolas. As gaiolas são o lugar onde as certezas moram.

É um engano pensar que os homens seriam livres se pudessem, que eles não são livres porque um estranho os engaiolou, que se as portas das gaiolas estivessem abertas eles voariam. A verdade é o oposto. Os homens preferem as gaiolas ao voo. São eles mesmos que constroem as gaiolas onde passarão as suas vidas."

R. Alves

E aqueles que foram vistos dançando foram julgados insanos por aqueles que não podiam escutar a música.

F. Nietzsche

Isto de querer ser exatamente o que a gente é, ainda vai nos levar além.

P. Leminsky

Dedico a minha mãe,

Pelo carinho e dedicação constantes,

E amor incondicional.

Agradecimentos

Agradeço ao meu orientador Fabio Gozzo primeiramente pela oportunidade de participar do seu grupo de pesquisa, pelos ensinamentos constantes, pelas discussões científicas de excelente nível não só no campo da química, mas incentivando sempre à busca pelo conhecimento, por ser sempre um apoio e agente motivador desse trabalho e, por fim, por continuar sendo uma fonte de inspiração intelectual.

Às pessoas que de maneira tão especial fizeram parte do meu cotidiano durante esse período, contribuindo direta ou indiretamente para o meu amadurecimento intelectual e pessoal, no laboratório e, muitas vezes, fora dele. Agradeço em espercial aos meus companheiros de lousa Mariana, Hugo, Luana e André que passaram continuamente a energia e o brio pelo conhecimento e reforçaram durante todo esse período a nossa amizade. A Tati, Lili, Dri, Paloma, Héctor, Alex, Mauro, Gisel e Marcel, agradeço pela boa convivência, pelo companheirismo e pelas experiências.

Ao Odirlei pelo cuidado e preocupação nos meus momentos mais difíceis. Pelo seu bom humor e carinho constantes e por fazer eu manifestar meu lado mais humano, me ensinando sempre com seu sorriso.

Ao Daniel (LNBio) que tão prontamente se disponibilizou a me ensinar a trabalhar com culturas de células e a expressar a STC1 além de participar ativamente em discussões a respeito do meu projeto. Obrigado pelas horas dedicadas e por um pouco da cultura que eu pude absorver daquilo que você me passou!

A Ângela por me ajudar no processo de adaptação ao LNBio e sempre se mostrar tão solicita, fazendo eu me sentir em casa. Obrigado também pelas conversas e amizade!

Ao Professor Jörg Kobarg e ao LNBio pela infraestrutura, materiais e disponibilidade para que eu pudesse desenvolver parte desse projeto.

Ao Instituto de Química da UNICAMP, pela infraestrutura e recursos humanos que colaboraram direta ou indiretamente para que esse trabalho fosse realizado.

Aos órgãos de fomento CNPq, Instituto Nacional de Ciência e Tecnologia em Bioanalítica e a Fapesp pela bolsa concedida.

Súmula Curricular

Dados pessoais

Nome Állan Jhonathan Ramos Ferrari
Nascimento 14/01/1991 - Londrina/PR - Brasil
CPF 058.257.269-00

Formação acadêmica

2013-2015 Mestrado em Química Orgânica
Universidade Estadual de Campinas, UNICAMP, Campinas, Brasil
Título: Caracterização Estrutural da Stanniocalcina-1 por Proteômica Estrutural
Orientador: Fabio Cesar Gozzo
Bolsista do(a): Fundação de Amparo à Pesquisa do Estado de São Paulo

2009 - 2012 Graduação em Química com Opção Tecnológica.
Universidade Estadual de Londrina, UEL, Londrina, Brasil
Título: Síntese do 4,6-di-O-acetil-2,3-didesoxi-D-eritro-hex-2-enono-1,5-lactona: um dienófilo para a investigação da reação de Diels-Alder como uma estratégia de síntese para Cardiopetalactona
Orientador: Fernando César de Macedo Júnior
Bolsista do(a): Conselho Nacional de Desenvolvimento Científico e Tecnológico

2009 Graduação interrompido(a) em Farmácia.
Centro Universitário Filadélfia, UNIFIL, Londrina, Brasil
Ano de interrupção: 2010

Produção Científica

Trabalhos apresentados em eventos

Ferrari, A., Cardoso, A.M., Kobarg, J., Balbuena, T.S. Gozzo, F. Cross-linking as a key experimental data in Stanniocalcin-1 structural modeling. 20th International Mass Spectrometry Conference, 2014, Genebra, Suíca.
Ferrari, A., Cardoso, A.M., Kobarg, J., Balbuena, T.S. Gozzo, F. Stanniocalcin-1 structural modeling. 2nd Meeting of the Brazilian Proteomics Society jointly with the 2nd

PanAmerican HUPO Meeting, 2014 – Búzios, RJ, Brazil.
Ferrari, A., Cardoso, A.M., Kobarg, J., Balbuena, T.S. Gozzo, F. Stanniocalcin-1 structural modeling based on chemical cross-linking coupled to mass spectrometry. V Congresso da Sociedade Brasileira de Espectrometria de Massas, 2013 – Campinas, SP.
Ferrari, A.. ; Wust, M. A. ; Butera, A. P. ; Fatima, A. ; Macedo JR, F. . Síntese da (-)-5β-hidroxigoniotalamina a partir da D-Glicose. XXVIII Semana de Química e V Jornada de pós-graduação em Química, 2012, Londrina.

Godoy, N. V.; Ribas, J. F.; **FERRARI, A**.; Abrao T. ; Santos, M. J. . Sorção de metais em solo e sedimento da área de mineração de carvão - avaliação de risco ambiental. 16o ENQA - Encontro Nacional de Química Analítica

Outras produções

06/2010-05/2011: **Iniciação Científica**

Título: Modelagem de sorção - dessorção de metais em solos e sedimentos - aspectos de transporte ambiental.

Agência Financiadora: PIBIC/CNPq

06/2011-02/2013: **Iniciação Científica**

Título: Síntese do 4,6-di-O-acetil-2,3-didesoxi-D-eritro-hex-2-enono-1,5-lactona: um dienófilo para a investigação da reação de Diels-Alder como uma estratégia de síntese para Cardiopetalactona

Agência Financiadora: PIBIC/CNPq

Resumo

CARACTERIZAÇÃO ESTRUTURAL DA STANNIOCALCINA-1 POR PROTEÔMICA ESTRUTURAL. A Stanniocalcina-1 (STC1) é um hormônio glicoproteico que apresenta padrão de expressão diferencial destacado em diversas patologias, notadamente em neoplasias, mas seus aspectos funcionais e estruturais são pouco explorados até o momento. Nesse sentido, a STC1 foi escolhida como alvo para a utilização de uma abordagem integrativa das técnicas que utilizam os reagentes de ligação cruzada, espectrometria de massas e modelagem molecular para a modelagem estrutural. A partir dos experimentos de ligação cruzada, foram obtidas 37 restrições de distância envolvendo espécies ligadas com DSS, sendo 11 destas espécies com N-terminal e uma restrição envolvendo a espécie dimérica, além das cinco ligações de dissulfeto já publicadas. Essas restrições foram utilizadas para a geração de modelos estruturais nas plataformas online I-Tasser e Quark e, localmente, mais de 100.000 modelos pelo protocolo *Ab Initio Relax* do software Rosetta em quatro diferentes condições iniciais de modelagem. O Rosetta apresentou maior eficiência na geração de modelos quando ausente arquivo de predição de estrutura secundária. As restrições de distância foram ferramenta discriminatória fundamental para a seleção de estruturas candidatas para a STC1. O agrupamento utilizando o parâmetro *global distante test (gdt)* das 1500 modelos de menor *score* que respeitavam todas as restrições identificou 22 estruturas representativas estruturalmente distintas. Essas estruturas representativas podem agora ser utilizadas em testes envolvendo substituição molecular nos dados de difração de raios-X.

Abstract

STRUCTURAL CHARACTERIZATION OF STANNIOCALCINA-1 BY STRUCTURAL PROTEOMICS. The Stanniocalcin-1 (STC1) is a glycoproteic hormone, which shows a differential expression pattern in a variety of pathologies, especially in neoplasia, but its functional and structural aspects have not been explored. Accordingly, the STC1 was chosen as a target to the use of an integrative approach including chemical cross-linking, mass spectrometry and molecular modeling. From cross-linking experiments,37 distance constrains were identified involving the DSS cross-linker, 11 of them in the N-terminus part of the protein and one involving the dimeric specie, in addition to five disulfide bonds already published. These constrains were used to generate structural models by I-Tasser and Quark online platforms and, locally, more than 100,000 models in the Ab Initio Relax protocol package present in the Rosetta software in four different modeling conditions. Rosetta was the most efficient in generating models when secondary structure prediction was absent. The distance constrains were found to be a key discriminatory tool for the selection of candidate structures for the STC1. For the 1,500 lowest score structures that satisfied the distance constrains, the clustering method employing the global distance test parameter (gdt) identified 22 structurally distinct representative structures. These representative structures can be used to in molecular replacement test to solve the X-ray diffraction data.

Sumário

Lista de Abreviaturas e Acrônimos

ALC Agente de Ligação Cruzada

CASP Critical Assessment of Techniques of Protein Structure Prediction

CD Espectroscopia de dicroísmo circular

CID Dissociação induzida por colisão

CV Volume da coluna

Da Dalton (1 Da = 1,661 x 10^{-24} g)

DDA Análise dependente de dados

DMF N,N-dimetilformamida

DRX Difração de Raios-X

DSS Suberato de N,N-disuccinimidila

DTT Ditiotreitol

ESI Ionização por Eletrospray

FPLC Fast Protein Liquid Chromatography

FRET Transferência de Energia Ressonante por Fluorescência

FTICR Ressonância Ciclotrônica de Íons com Transformada de Fourier

GDT Global Distance Test

HDX Troca Hidrogênio/Deutério

HPLC Cromatrografia Líquida de alta eficiência

HT His-Tag

IEX Tampão para cromatografia por troca iônica

IMS Espectrometria de Mobilidade Iônica

LC Cromatrografia Líquida

LIT IonTrap Linear

LLA Leucemia Linfóide Aguda

MALDI Ionização por Dessorção a Laser auxiliada por Matriz

MaxSub Maximal subset

MD Dinâmica Molecular

MES Ácido 2-(N-morfolino)-etanosulfônico

MS	Espectrometria de Massas
MS^n	Espectrometria de massas sequencial
NHS	N-hidroxisuccinimida
NMWL	Peso molecular limite nominal
PCR	Polimerase Chain Reaction
PDB	Protein Data Bank
Q	Analisador de massas do tipo quadrupolo
RMN	Ressonância Magnética Nuclear
SAXS	Espalhamento de raios-X a baixo ângulo
SIM	Spectrum Identification Machine
SizeEx	Tampão para cromatrografia por exclusão
STC	Stanniocalcina
SUMO	Small Ubiquitin-Like Modifier
TOF	Analisador de massas do tipo tempo de voo
vdW	Van der Walls

Índice de Tabelas

Índice de Figuras

1 . Introdução

1.1 Estrutura de Proteínas

Proteínas são macromoléculas presentes em todos os sistemas biológicos e possuem especial interesse por serem as grandes efetoras na dinâmica desses sistemas. Essas entidades moleculares são compostas por aminoácidos ligados covalentemente por ligações peptídicas e se apresentam em uma ampla faixa de formas e tamanho, variando de poucos kDa à alguns MDa[1] (**Figura 1**).

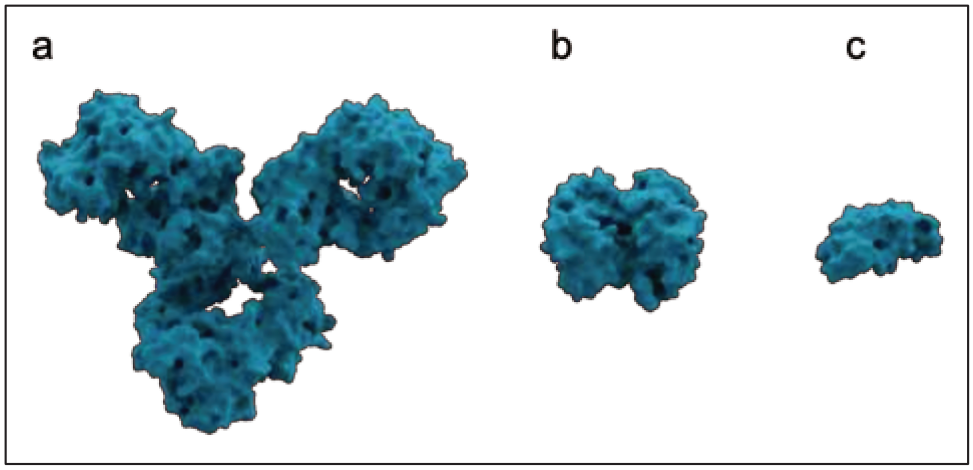


Figura 1: Representação da diversidade de tamanhos e formas de proteínas. Em (A) estrutura da imunoglobulina (IgG), anticorpo envolvido em respostas imunes secundárias; (B) hemoglobina, envolvida no transporte de oxigênio; e (C) insulina, hormônio envolvido na redução da glicemia.

Interações de hidrogênio locais entre resíduos de aminoácidos sequenciais dão origem a estruturas regulares conhecidas como alfas-hélices e folhas-beta, elementos de estrutura secundária. A cadeia polipeptídica, produto da síntese proteica, além de assumir localmente estas estruturas, enovela-se de forma a globalmente esconder os resíduos hidrofóbicos e expor os resíduos hidrofílicos, maximizando assim as interações com o solvente e dando origem a estrutura nativa funcional, conhecida como estrutura terciária. Cadeias polipeptídicas distintas podem ainda formar uma estrutura funcional supramolecular, correspondendo a um nível de estrutura quaternária[1] (**Figura 2**).

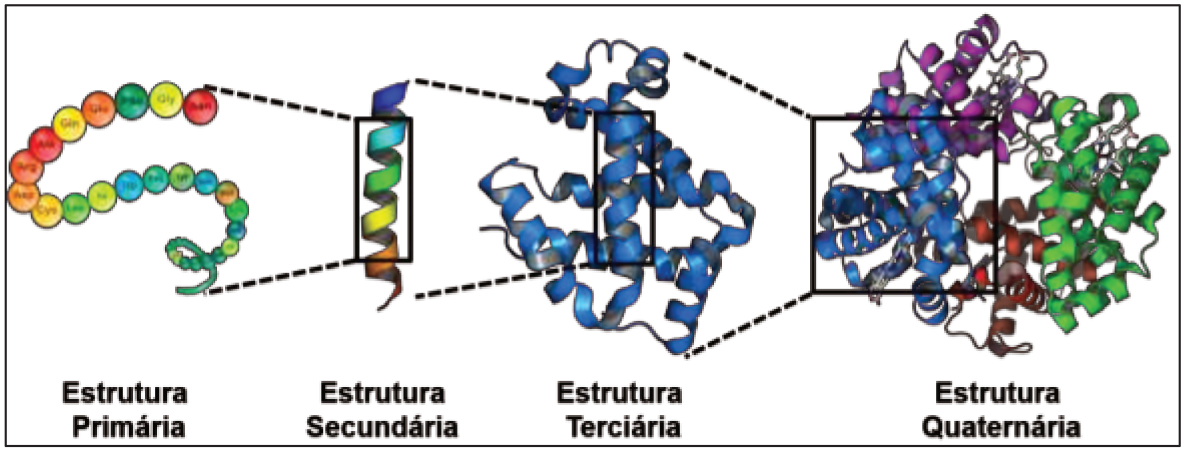


Figura 2: Esquema mostrando a divisão sistemática entre os níveis estruturais de uma proteína. A estrutura primária compõe a sequencias de aminoácidos ligados por ligações peptídicas; esses aminoácidos dão origem a estruturas regulares mantidas por ligações de hidrogênio: alfa-hélices e folhas-beta; a estrutura terciária é mantida por interações não locais e é o produto do enovelamento proteico; estruturas supramoleculares podem ser formadas por interações não covalentes entre subunidades proteicas dando origem à estrutura quaternária.

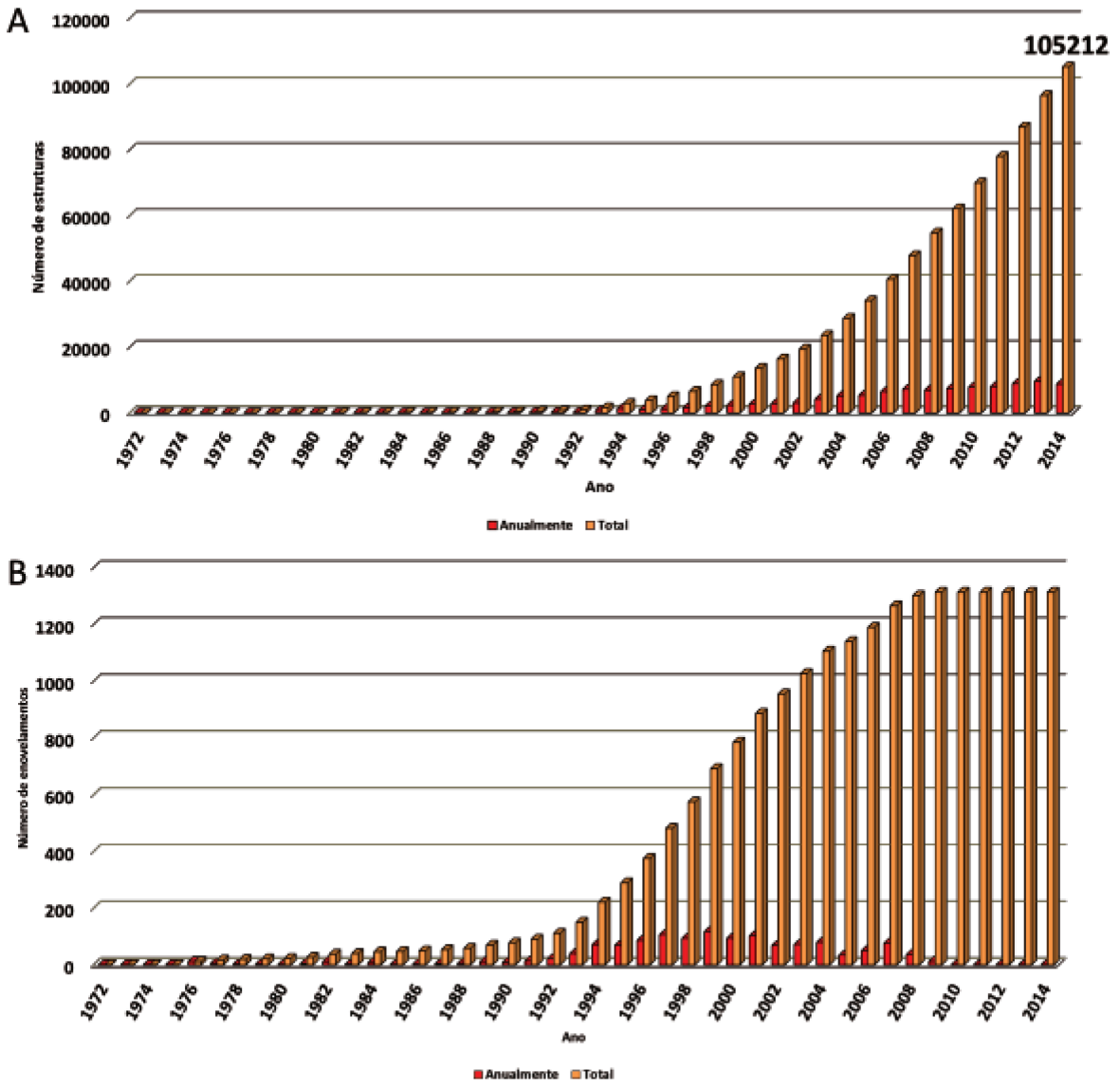
A estrutura nativa das proteínas está diretamente ligada a sua estabilidade e atividade biológica, portanto, com exceção das proteínas intrinsicamente desordenadas, alterações em suas estruturas podem provocar a perda de sua função biológica[1,2].

Novas tecnologias têm sido desenvolvidas para a análise em larga escala da função de uma proteína, em que as descrições geralmente envolvem apenas alguns dos vários aspectos funcionais, tais como localização, interação e modificações, mas raramente com a descrição total de suas propriedades. Detalhes a respeito dos mecanismos moleculares envolvidos em determinado processo e, assim, na função de determinada proteína, requerem estudos muito mais dispendiosos e, necessariamente, passam pela avaliação estrutural dos componentes do sistema de interesse.

1.2 Métodos de Determinação Estrutural

As técnicas de determinação estrutural de proteínas e complexos proteicos podem ser classificadas de acordo com o grau de resolução que elas oferecem. As técnicas de baixa resolução são constituídas em geral pelas espectroscopias, tais como: Dicroísmo Circular (CD), Emissão de Fluorescência, Transferência de Energia

Ressonante Förster (FRET) entre outras. Técnicas de ultra-centrifugação também constituem outro importante componente das técnicas de baixa resolução. Essas técnicas em geral são de mais fácil acesso e relativamente simples e rápidas, porém a informação estrutural normalmente é muito restrita. Técnicas com maior grau de resolução englobam microscopias e o espalhamento de raios-X a baixos ângulos (SAXS), e normalmente essas técnicas são mai



s restritas devido à instrumentação mais complexa e específica. Essas técnicas fornecem informações relativas à estequiometria, posição e orientação dos componentes do sistema.

A determinação de proteínas com resolução atômica, ou alta resolução, pode atualmente ser obtidas por duas técnicas, sendo elas a Ressonância Magnética Nuclear (RMN) e a Difração de Raios-X (DRX), cujas estruturas resolvidas são

geralmente depositadas no *Protein Data Bank* (PDB). A análise da **Figura 3A** permite verificar que do número de estruturas no PDB cresce exponencialmente, chegando hoje a mais de 100.000 estruturas. Destas, aproximadamente 89% foram resolvidas por DRX enquanto que em torno de 10% foram resolvidas por RMN. Além disso, é interessante notar na **Figura 3B** que desde 2009 não se determina nenhuma estrutura de proteína com um novo enovelamento básico, mostrando que, aparentemente, existe um número limitado de topologias que podem ser determinadas para essas macromoléculas pelas técnicas de alta resolução.

Dentro deste contexto, torna-se possível que a determinação estrutural (dentro de certos limites de resolução) possa ser realizada pela determinação dos enovelamentos já conhecidos que compõe a estrutura de uma determinada proteína.

Figura 3: Estatísticas encontradas no PDB. O gráfico (A) indica que o número de estruturas resolvidas e depositadas cresce exponencialmente, enquanto que o número de enovelamentos, mostrado em (B) se mantém constante desde 2009.

Apesar do grande sucesso das técnicas de alta resolução, muitas proteínas não são passíveis de análise por essas técnicas, pois em geral há a necessidade de uma alta concentração e em elevado grau de pureza. Ainda no caso de RMN, o tamanho está limitado a aproximadamente 30 kDa e a proteínas que sejam estáveis em tampões que não interfiram na análise por dias ou até semanas. Uma estatística do *New York Structural Genomics Research Consortium* (NYSGR) (**Figura 4**) revela que para a DRX a grande maioria dos alvos não chegam a etapa de análise. A obtenção de monocristais é o grande gargalo da técnica, sendo que a taxa de sucesso em relação ao número de proteínas purificadas é próxima de 10% e quando comparado com o número de proteínas clonadas essa taxa cai para próximo de 2%.

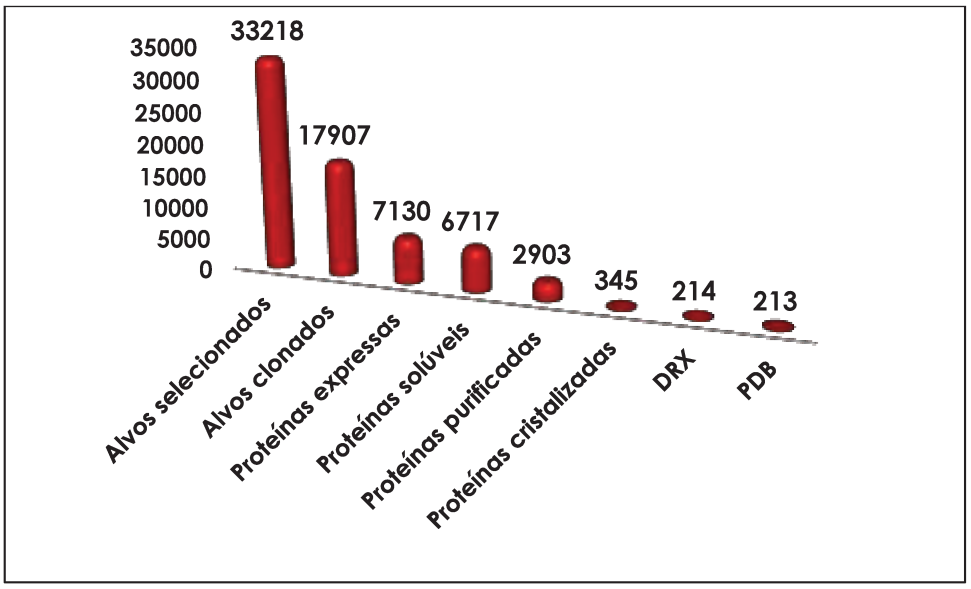


Figura 4: Estatísticas do NYSGR indicando que aproximadamente 2% dos alvos clonados possuem sua estrutura resolvida e depositada do PDB pela técnica de DRX.

Uma análise comparativa entre os dados do PDB com o número de sequencias de proteínas derivadas de análises genômicas depositadas no SwissProt (http://web.expasy.org/docs/relnotes/relstat.html) presentes na **Figura 5A** revela que, mesmo com o aumento exponencial do número de estruturas, as técnicas de alta resolução não são suficientes para acompanhar a velocidade com que se tem acesso as informações advindas do sequenciamento de genomas. São mais de 540 mil sequencias depositadas contra pouco mais de 105 mil estruturas no PDB, correspondendo a aproximadamente 19% de estruturas em relação ao total de sequencias. Tal divergência se torna ainda mais relevante quando analisa-se um banco de sequencias mais completo, como o TrEMBL (http://www.ebi.ac.uk/uniprot/TrEMBLstats) (**Figura 5B**). A porcentagem de estruturas resolvidas em relação ao número de sequencias depositadas cai para pouco mais do que 0,11%. Nesse sentido, o grande desafio da era pós-genômica está em se assinalar a estrutura e a função dos produtos dos genes de todos os organismos.

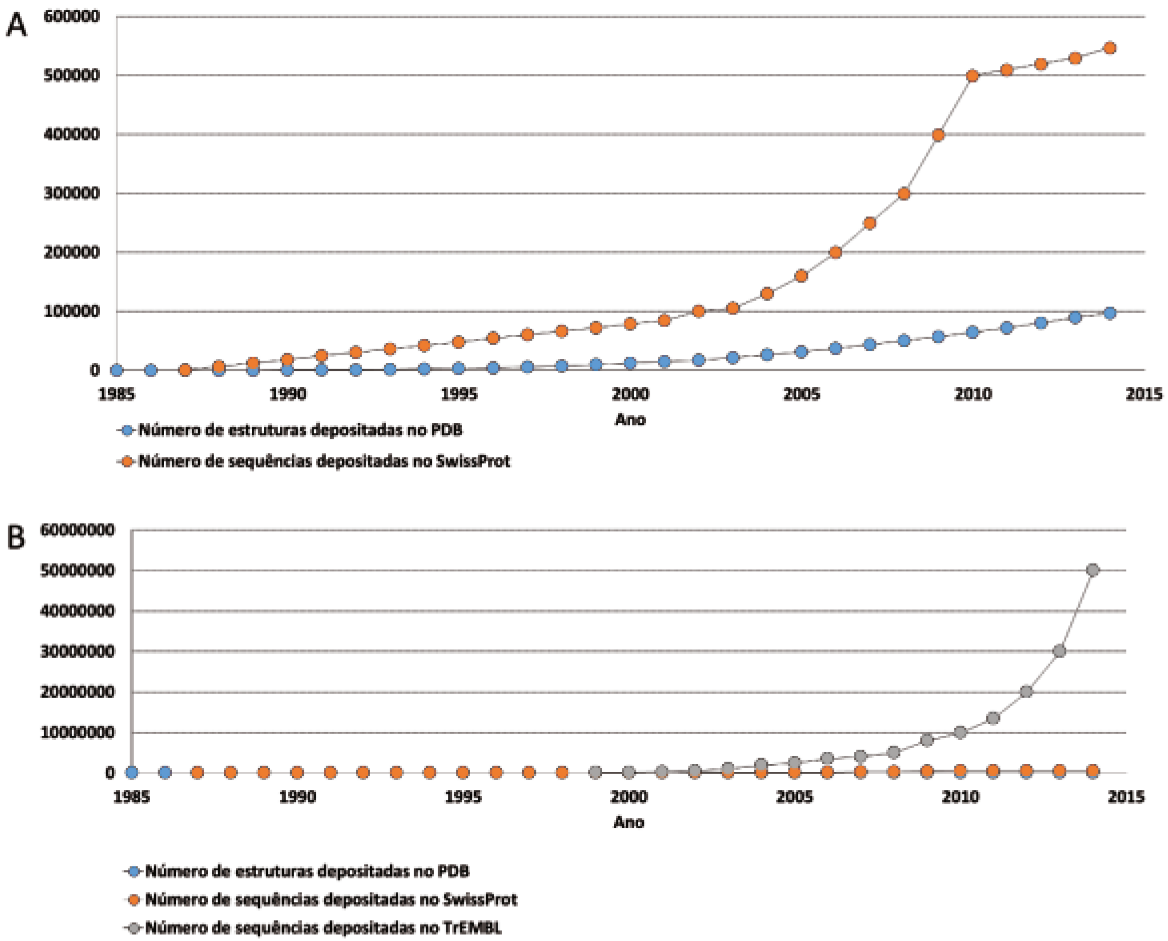


Figura 5: Estatísticas de bando de dados de sequencias genômicas em relação aos dados do PDB. Em (A) é indicado o comportamento exponencial do número de sequencias de proteínas depositadas do SwissProt, enquanto que ao gráfico (B) é adicionado o número de sequencias de proteínas do TrEMBL indicando que as técnicas atuais de determinação de estrutura de proteínas de alta resolução não são capazes de acompanhar a velocidade dos sequenciamentos genômicos.

Devido a este fato, diversos grupos têm investido em técnicas computacionais e algoritmos sofisticados com o objetivo de prever a estrutura terciária de proteínas com base em sua sequência primária. A modelagem de proteínas é uma ferramenta essencial no sentido de preencher a distância que existe entre o número de estruturas determinadas experimentalmente e o número de proteínas previstas pelas análises genômicas. Esses programas de predição de estruturas são avaliados bianualmente pelo CASP (*Critical Assessment of Techniques for Protein Structure Prediction*) em uma competição em que é fornecida uma sequência de uma proteína com estrutura determinada, porém ainda não publicada, para que os grupos, dentro de um

determinado período, possam obter modelos que são então avaliados quanto a sua precisão em relação à estrutura experimental de alta resolução.

1.3 Predição de Estruturas de Proteínas

A predição de estruturas de proteínas a partir da sequência de aminoácidos que a compõe ainda é um problema não resolvido após cinco décadas de esforços. Se uma proteína alvo possui um homólogo já determinado, a tarefa é relativamente mais simples e modelos de alta resolução podem ser construído pela cópia da estrutura básica da estrutura resolvida. Entretanto, tal procedimento não responde a questão de como e o porquê a proteína adota aquela determinada estrutura. Esse procedimento é chamado de modelagem baseada em modelo (*template-based modelling)*[3,4]. Ao contrário da modelagem baseada em modelos, se estruturas homólogas não existem, ou existem mas não podem ser identificadas por similaridade de sequências, modelos devem ser construídos sem o uso de dados experimentais. Esse procedimento, chamado de modelagem *ab initio*[5,6] (também chamada abordagem *de novo*, baseado em física ou modelagem livre), é essencial para uma completa solução do problema de predição de estrutura de proteínas; também pode ajudar a compreender qual o princípio físico-químico de como as proteínas se enovelam na natureza. Atualmente, a precisão da modelagem *ab initio* é baixa e o seu sucesso é limitado a pequenas proteínas (< 100 resíduos).

Tipicamente, a modelagem *ab initio* conduz uma busca conformacional através da avaliação de uma função de energia. Esse procedimento normalmente gera um número de possíveis conformações, e os modelos finais são selecionados dentre eles. Portanto, uma modelagem *ab initio* de sucesso depende de três fatores: (1) uma função de energia precisa com a qual a estrutura nativa da proteína corresponda ao estado termodinâmico mais estável, comparado com as outras conformações avaliadas; (2) um método eficiente de busca conformacional que possa rapidamente identificar estados de baixa energia; (3) uma ferramenta seleção de modelos “nativos” dentre as estruturas geradas.

As funções de energia podem ser classificadas em dois grupos distindos: (a) funções de energia baseadas em física (*physics-based energy function*) e (b) funções de energia baseadas em estatísticas (*knowledge-based energy function*).

Em um método em que a função de energia consiste estritamente em física, interações entre os átomos devem ser baseadas em mecânica quântica com apenas alguns parâmetros fundamentais como a carga do elétron e a constante de Planck; todos os átomos devem ser descritos pelos seus tipos atômicos onde somente o número de elétrons é relevante[7]. Entretanto, não há nenhuma tentativa relevante de se iniciar a predição de estruturas a partir de mecânica quântica, mesmo de pequenas proteínas, simplesmente porque os recursos computacionais para tais cálculos estão muito distantes dos atualmente disponíveis. Sem os tratamentos da mecânica quântica, um início prático para a modelagem *ab initio* é utilizar um campo de forças com um grande número de tipos atômicos selecionados; em cada tipo atômico, as propriedades físicas e químicas dos átomos são adquiridas de cálculos do empacotamento cristalino ou da teoria da mecânica quântica para pequenas moléculas. Exemplos de tais campos de força incluem o AMBER[8], CHARMM[9] e OPLS[10]. Estes potenciais contêm termos associados com comprimento de ligação, ângulos, interações de van der Waals e eletrostáticas. A maior diferença entre eles está na seleção dos tipos atômicos e nos parâmetros de interação.

As funções de energia com base em estatística possuem como referência termos empíricos derivados de estruturas resolvidas e depositadas no PDB[11]. Esses termos podem ser divididos entre aqueles genéricos e independentes de sequência, tais como ligação de hidrogênio e rigidez da cadeia polipeptídica ou aqueles derivados de tendências observadas que são dependentes de sequência, como potencial derivado do contato entre pares de resíduo, potencial derivado da dependência de distância entre átomos e tendências relacionadas a estrutura secundária[12,13]. Uma das formas de aplicação desse tipo de abordagem incluem a utilização de uma biblioteca de fragmentos de estrutura secundária, obtidas pelo alinhamento diretamente das estruturas resolvidas que reduzem significativamente a entropia da busca conformacional[14].

Atualmente, I-Tasser[15–17], Quark[18] e Rosetta[19–21], alguns dos programas melhores ranqueados pelo CASP, utilizam uma abordagem muito parecida baseando-se essencialmente na construção de uma biblioteca de fragmentos de estrutura secundária, associadas a uma função de energia baseada em estatística que avalia as montagens dos modelos. O Rosetta ainda utiliza na etapa de refinamento uma

função baseada em física que avalia diferentes rotâmeros para a cadeia lateral, etapa mais dispendiosa computacionalmente.

Outra possibilidade bastante atrativa é a utilização de dados experimentais como guias destes programas, visto que pode levar a construção de modelos mais refinados ou ser uma ferramenta de seleção dos modelos gerados.

Apesar dos avanços nesta área, a resolução da estrutura por métodos computacionais ainda apresenta uma exatidão relativamente baixa e recentemente o acoplamento de métodos computacionais com dados experimentais tem atraído bastante atenção. Exemplos desse tipo de abordagem incluem o SAXSTER[22] e o Modeller IMP[23]. Dentre os dados experimentais mais utilizados para guiar a modelagem da estrutura, encontram-se os dados de SAXS, microscopia, RMN e MS, tendo este último se apresentado teoricamente como um dos mais valiosos dados experimentais para tal fim.[24]

1.4 Espectrometria de Massas no Estudo de Proteínas

A espectrometria de massas (MS) consiste no estudo de íons na fase gasosa, sendo a caracterização estrutural de compostos uma de suas principais aplicações. O interesse na aplicação dessa técnica para a análise de proteínas e outras biomoléculas existe desde a década de 70. Entretanto, a dificuldade envolvida na geração de íons de macromoléculas em fase gasosa pelos métodos até então disponíveis era uma das principais limitações da técnica[25,26]. O advento das técnicas de ionização suaves denominadas ESI (*Eletrospray Ionization)*[27], por J. B. Fenn e MALDI (Matrix-Assisted Laser Desorption Ionization)[28], por M. Karas e F. Hillenkamp, no final da década de 80, possibilitou a análise de biomoléculas como peptídeos, proteínas e ácidos nucléicos, sem fragmentação ou derivatização[27]. O desenvolvimento dessas técnicas de ionização impulsionou uma grande evolução dos analisadores de massas associados nos anos subsequentes, de tal maneira que a partir da década de 90 tornaram-se disponíveis instrumentos comerciais com capacidade de experimentos de MS sequencial (MS^n) e diversas geometrias, como quadrupolo tempo de vôo (Q-TOF), tempo de vôo – tempo de vôo (TOF-TOF), ion trap linear (LIT), ressonância ciclotrônica de íons com transformada de Fourier (FTICR) e, mais recentemente, Orbitrap[29,30]. Desde então, o uso da espectrometria

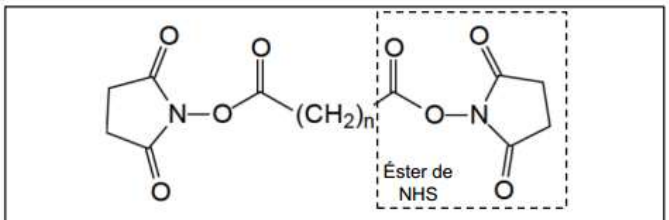
de massas na análise de proteínas evoluiu extensivamente, a ponto de se tornar rotineiro no sequenciamento e identificação de proteínas, determinação da massa molecular de proteínas e complexos proteicos intactos, determinação de parceiros de interação, identificação e localização de modificações pós-traducionais e quantificação absoluta e relativa de proteínas[31–37]. Em particular, o desenvolvimento de metodologias envolvendo cromatografia líquida de alta eficiência (HPLC) acoplada a MS para análises de misturas complexas de peptídeos advindos de digestão enzimática, contendo uma enorme diversidade de proteínas, aliada ao crescimento das ferramentas de bioinformática para busca em banco de dados possibilitou estudos de proteômica em larga escala, hoje denominados de proteômica *shotgun*.

O interesse na ampla aplicação de MS para a caracterização de proteínas se deve às vantagens intrínsecas da técnica, como alta sensibilidade (permitindo análise de amostras contendo picomols de peptídeos), rapidez, versatilidade, facilidade de operação e alta confiabilidade dos resultados. Como demonstrado acima, essas características fazem com que MS seja hoje uma técnica bastante consolidada para caracterizações envolvendo a estrutura primária de proteínas. No entanto, o emprego de MS na análise de estruturas superiores (terciárias e quaternárias) de proteínas ainda é restrito, apesar das inúmeras vantagens potenciais que poderia apresentar nessa área. Nesse âmbito, metodologias envolvendo análise por MS têm se destacado recentemente, com um número crescente de aplicações: mapeamento radicalar (*Footprinting*), Mobiblidade Iônica (IMS), troca Hidrogênio/Deutério (HDX) e Ligação cruzada (*Cross-linking*).

1.5 Ligação Cruzada Acoplada a Espectrometria de Massas para o Estudo Estrutural de Proteínas

O fenômeno de ligação cruzada (*cross-linking)* compreende a união de duas espécies através de uma ligação covalente, normalmente se empregando um agente de ligação cruzada (ALC). Diferentes classes químicas podem estar envolvidas em experimentos de ligação cruzada, variando de proteínas e ácidos nucléicos a partículas sólidas[38]. Ligações de dissulfeto presentes em proteínas são exemplo de ligações cruzadas que ocorrem naturalmente.

Os ALCs são compostos orgânicos multifuncionais, contendo em geral dois ou três grupos reativos, unidos por uma cadeia espaçadora de comprimento variável. Quando em contato com proteínas em solução, os ALC são capazes de se ligar covalentemente a cadeias laterais dos resíduos de aminoácidos, de acordo com suas especificidades, unindo cadeias laterais que estejam espacialmente separadas, no máximo, pelo comprimento da cadeia espaçadora[39]. Os grupos reativos podem ser idênticos (homofuncionais) ou distintos (heterofuncionais), permitindo versatilidade na especificidade de cada reagente. Dentre os ALCs disponíveis atualmente, derivados de ésteres de N-hidroxisuccinimida (NHS) são os mais utilizados (**Figura 6**), devido aos rendimentos relativamente altos das reações e, principalmente, à facilidade de sua obtenção, bastando para isso reagir um haleto de acila com um NHS. Estes reagentes são reativos preferencialmente frente a aminas primárias, podendo reagir também com grupos álcoois e tióis. Outros ALCs são também empregados com outras especificidades: carbodiimidas frente a ácidos carboxílicos; isocianatos frente a álcoois; e maleimidas frente a tióis[40], por exemplo.



n	Nome	Distância (Å)
3	Glutarato de Dissuccinimidila (DSG)	7,7
5	Pimelato de Dissuccinimidila	9,5
6	Suberato de Dissuccinimidila (DSS)	11,4
8	Sebacato de Dissuccinimidila (DSSeb)	13,0

Figura 6: Série de agentes de ligação cruzada homólogos derivados de NHS com diferentes tamanhos de cadeia alifática.

As informações advindas dos experimentos de ligação cruzada acoplado a MS podem ser utilizadas para a obtenção de informações estruturais de proteínas. Tais informações são provenientes do tamanho da cadeia espaçadora do ALC utilizada, que limita os dois resíduos de aminoácidos envolvidos a uma determinada distância. Em 2000, foi proposto pela primeira vez que se o número de restrições de distância

fosse maior que a N/10 (onde N é o número de resíduos da proteína), seria possível se determinar o tipo de enovelamento por meio de métodos computacionais[41].

A MS permite a análise de produtos envolvidos na ligação cruzada, aproveitando-se de todas as vantagens intrínsecas da técnica. Isso é realizado submetendo a proteína modificada pelo ALC à proteólise enzimática e identificando-se os peptídeos modificados por LC-MS e LC-MS/MS, sendo a mistura de peptídeos submetidas à separação cromatográfica seguida de análise por MS de forma contínua. Os íons de peptídeos são automaticamente submetidos a experimentos de MS/MS durante a corrida de LC, permitindo identificá-los e localizar as modificações causadas pelo ALC. Essa abordagem é semelhante aos métodos de proteômica *shotgun* empregados em análises de misturas complexas de proteínas (**Figura 7**).

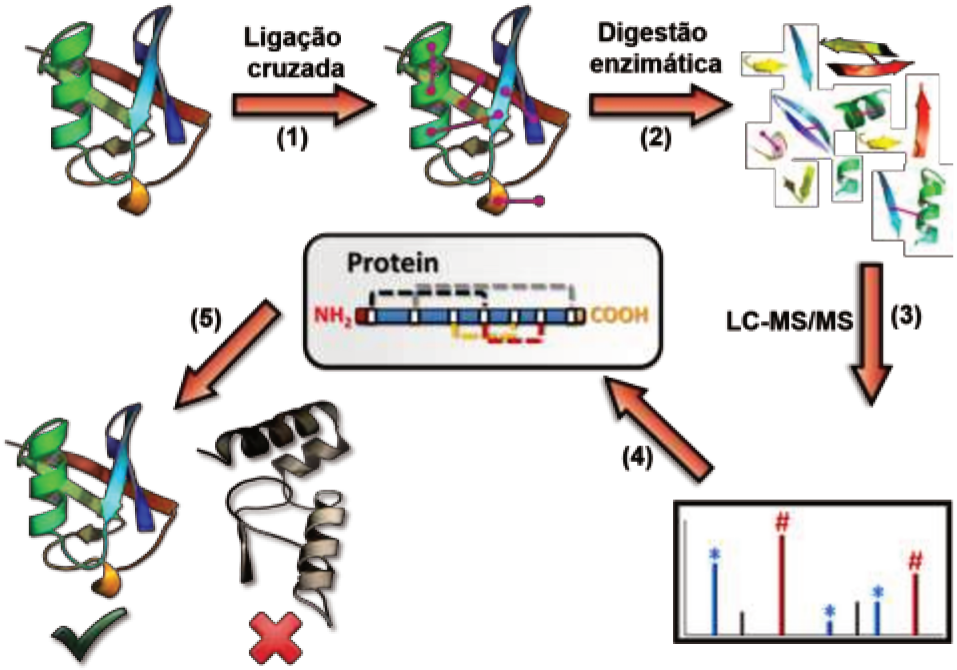


Figura 7: Representação esquemática de um experimento de ligação cruzada acoplado a espectrometria de massas. (1) Reação da proteína com o ALC; (2) Proteólise enzimática dando origem às diversas espécies de peptídeos, convencionais e modificadas; (3) Análise da mistura por LC-MS/MS; (4) Agrupamento das restrições de distância obtidas experimentalmente em um mapa de restrições de distância; (5) Utilização do conjunto de restrições para predição de estruturas ou seleção de modelos compatíveis.

Os produtos de proteólise enzimática derivados da reação das proteínas com o ALC podem ser classificados em três tipos (**Figura 8**): 1) espécies inter-

peptídicas, em que a cadeia espaçadora conecta dois peptídeos que estariam separados devido à proteólise enzimática, e fornecem informações a respeito de resíduos próximos na estrutura terciária ainda que estejam distantes na sequência primária; 2) espécies intra-peptídicas, em que a cadeia espaçadora conecta dois resíduos de aminoácidos de um mesmo peptídeo gerado na proteólise enzimática, e informam que resíduos próximos na sequência primária também se encontram dentro da restrição da cadeia espaçadora na estrutura terciária; e 3) espécies parcialmente hidrolisadas ou *dead end*, em que uma das porções do ALC reagiu com uma cadeia lateral da proteína e a outra porção sofreu hidrólise, que trazem informações a respeito da acessibilidade ao solvente desse resíduo.

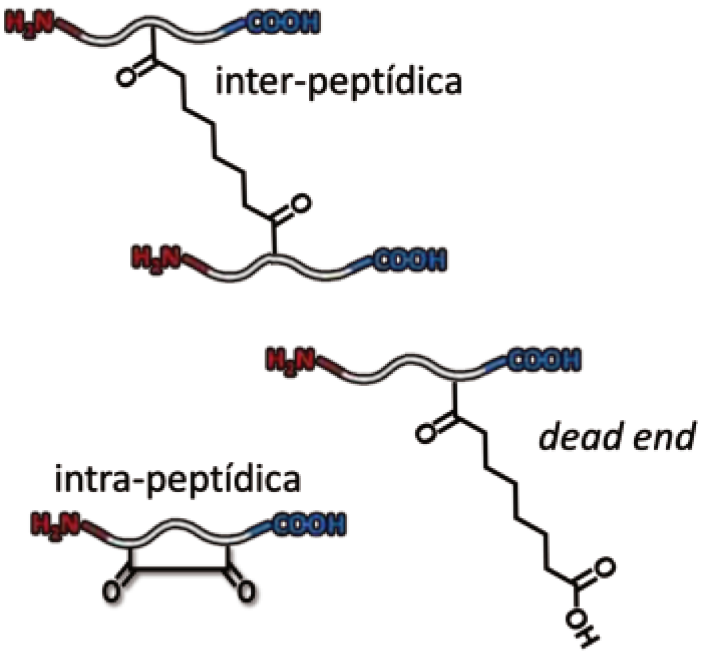


Figura 8: Representação das espécies modificadas encontradas em um experimento de ligação cruzada.

Apesar de muito promissora na análise de complexos de proteínas e estruturas de proteínas, a metodologia de ligação cruzada acoplada a MS apresenta alguns desafios, como detecção e identificação dos peptídeos modificados pelo ALC, pois estes se apresentam em quantidades subestequiométricas quando comparados com os peptídeos não modificados.[39,42] Entre as estratégias que tem sido utilizadas encontram-se o uso de ALCs marcados isotopicamente[43,44], que torna a síntese do ALC mais cara e a análise mais complexa; o uso de ALC modificados com grupos específicos visando purificação por cromatografia de afinidade dos peptídeos

modificados, o que provoca perda da amostra por extensa manipulação[45,46], e o uso de ALC cliváveis visando facilitar a detecção dos peptídeos modificados[47], que necessita de instrumentos capazes de realizar experimentos de espectros de íons fragmentos do tipo multi-estágio (MS^n).

Nosso grupo tem trabalhado com essa metodologia obtendo informações em diversos níveis estruturais, como na obtenção de informação a respeito da estrutura secundária[48], informações sobre complexos proteicos[49,50] e mudanças conformacionais de proteínas[51]. Além disso, estudos fundamentais sobre fragmentação dessas espécies modificadas com derivados de NHS permitiram a identificação de íons marcadores nos espectros de MS/MS[52]. Esses marcadores facilitam a identificação dos peptídeos modificados e tem sido utilizados em um software recentemente desenvolvido em nosso grupo (SIM – *Spectrum Identification Machine)* para a identificação de peptídeos contendo ligação cruzada.[53] A avaliação teórica de diversas técnicas experimentais para auxiliar na determinação de complexos proteicos[24] através de *docking* molecular indicam que a utilização das informações de restrição de distância na modelagem dessas estruturas possuem um efeito muito significativo, no mínimo iguais, mas muitas vezes superior a SAXS, RMN e microscopias 2D e 3D (EM2D e 3D, respectivamente) (**Tabela 1**).

Tabela 1 – Taxa de sucesso do docking molecular integrativo utilizando diferentes filtros experimentais. (A técnica de ligação cruzada é representada na Tabela como CXMS)

Top N	Standard docking	Standard docking EM cases	SAXS	EM2D	EM3D	NMR-RTC	CXMS
1	10%	7%	22%	33%	33%	18%	36%
10	24%	19%	51%	82%	79%	47%	65%
100	49%	26%	77%	89%	89%	76%	87%
Case no.	176	27	176	27	27	176	138

Apesar da técnica de ligação cruzada ter mostrado um grande potencial na obtenção de informações em diversos níveis estruturais de proteínas, ainda não

existe um estudo sistemático a respeito do potencial da técnica na determinação de estrutura terciária de proteínas.

1.6 A Stanniocalcina-1

As Stanniocalcinas (STC) são proteínas pertencentes a uma pequena família de hormônios glicoproteicos secretados, composta pelas proteínas STC1 e STC2 cujas sequências de aminoácidos são altamente conservadas desde vertebrados aquáticos até os terrestres[54,55].

Inicialmente descoberta em peixes ósseos associada à homeostase de cálcio[56], fica cada vez mais claro que as STCs tenham expandido seus papeis em mamíferos, por apresentarem padrão de expressão abrangente em diversos tecidos adultos[54,57,58] e também apresentarem expressão diferencial durante embriogênese[57] e diversos tipos de câncer[59–62]. Apesar disso, sua caracterização funcional ainda é muito pequena.

Recentemente o grupo do Prof. Dr. Jörg Kobarg (IB-Unicamp) interessado em estudar Leucemia Linfóide Aguda (LLA) identificou o gene da STC1 como um potencial biomarcador para o seu ambiente tumoral (**Figura 9**) e caracterizou funcionalmente a proteína como uma SUMO E3 ligase[63]. A sumoilação é uma modificação pós-traducional envolvida em vários processos celulares, tais como: transporte, regulação transcricional, apoptose, estabilidade proteica, resposta a estresse e ciclo celular[64,65].

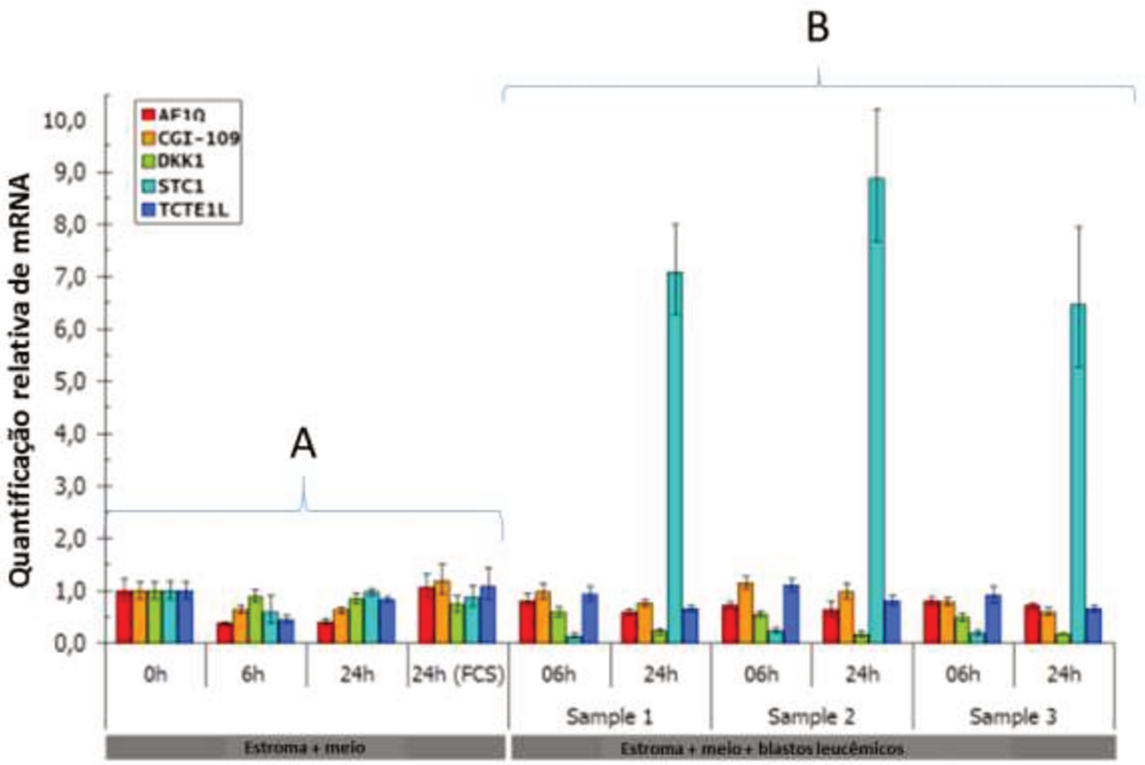


Figura 9: Análise da expressão de genes candidatos por PCR Quantitativo em tempo real (qPCR). (A) Células do estroma da medula óssea foram crescidas até alcançarem confluência e a expressão do mRNA de diferentes genes candidatos tanto na presença quanto na ausência de soro bovino foram analisados após 6h e 24h; (B) o mesmo procedimento foi realizado para três amostras estimuladas com blastos leucêmicos de três diferentes pacientes. A STC1 apresenta após 24h um grande aumento dos níveis de expressão [adaptado da referência 63].

A STC1 de humanos e ratos possui 247 resíduos de aminoácidos[66], cujos 17 primeiros compõe um peptídeo sinal, responsável pela capacidade dessa proteína ser secretada para o meio extracelular. Os 204 primeiros resíduos apresentam 92% de similaridade de sequência da STC de salmão, incluindo um sítio de N-glicosilação[59]. Além disso, possui um padrão de formação de cinco ligações de dissulfeto extremamente conservada conservadas (**Figura 10**). Entretanto, a porção C-terminal é pouco conservada, o que pode sugerir que a principal atividade biológica da STC1 deva estar relacionada com sua porção N-terminal[56,67].

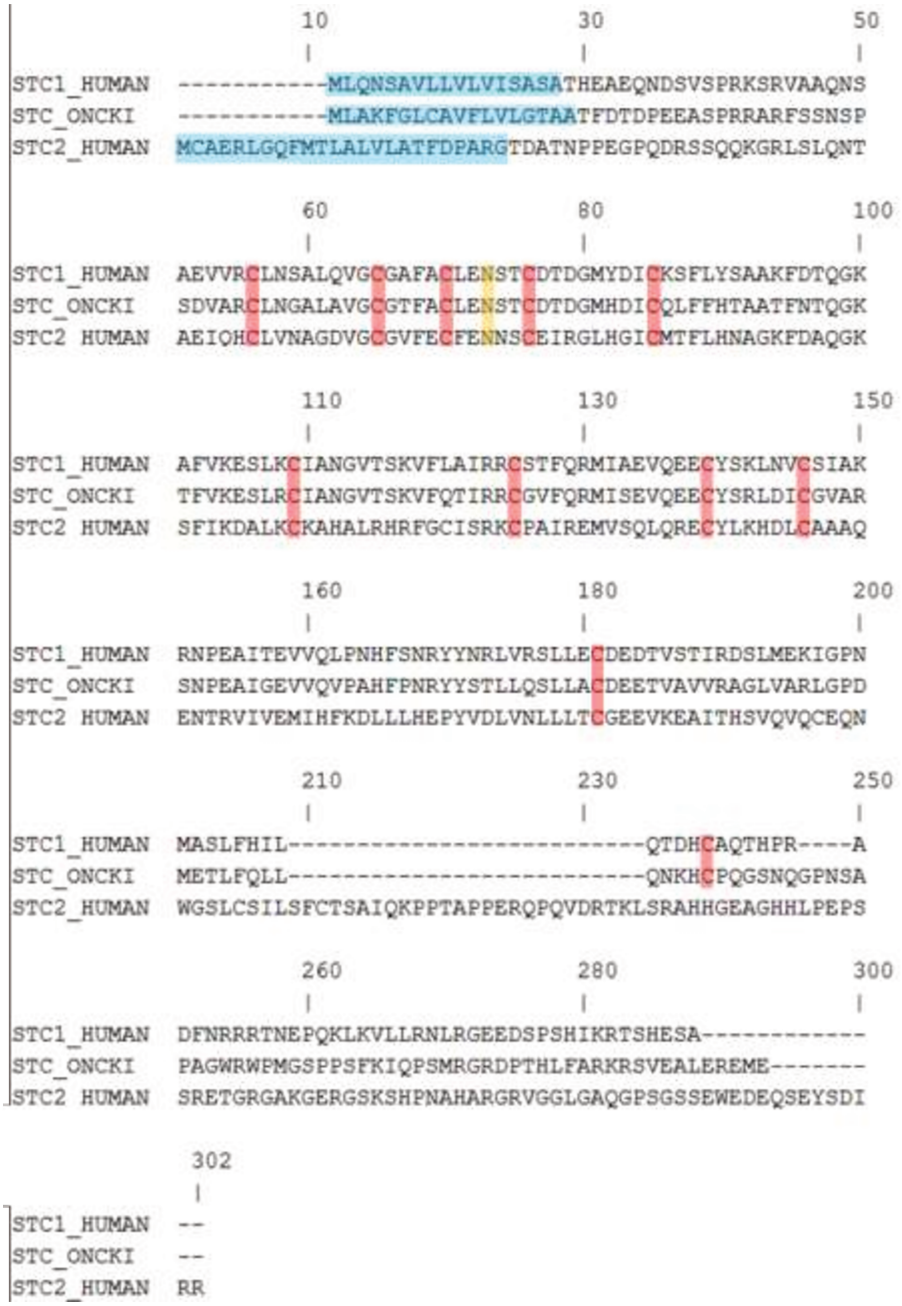


Figura 10: Alinhamento entre sequências de STC-1 e -2 humanas e a STC de truta arco-íris (Oncorhynchus mykiss, família Salmonidae). Marcações em azul: peptídeo sinal, vermelho: cisteínas envolvidas em ligações de dissulfeto e em laranja: sitio de N-glicosilação.

Estudos estruturais de baixa resolução[68] revelaram que a STC1 é uma proteína estruturada composta basicamente por alfa-hélices e, em solução, se apresenta em um formato alongado (**Figura 11**). Além disso, o grupo do Prof. Kobarg conseguiu cristalizar a STC1 e obter seu padrão de difração, porém não conseguiu resolver a

sua estrutura.

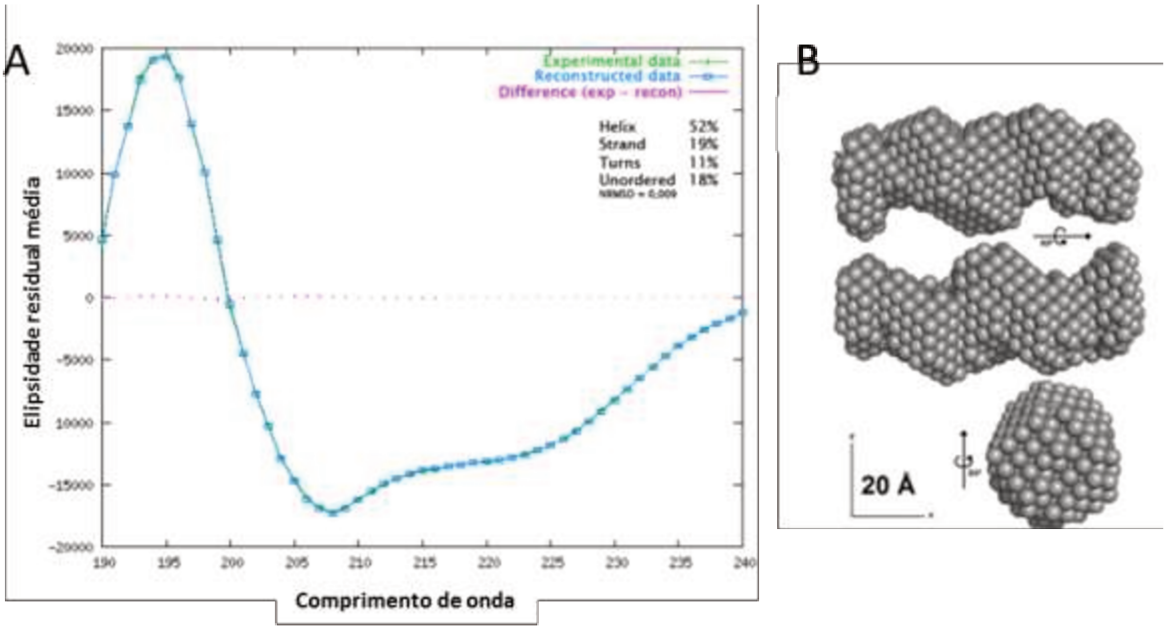


Figura 11: Estudos revelam que a STC1 é uma proteína estruturada. (A) Espectro de dicroísmo circular mostrando que a STC1 é formada, em sua porção estruturada, essencialmente por alfa-hélices. (B) Estudos de baixa resolução envolvendo SAXS revelam em um modelo ab initio que a STC1 possui formato alongado em solução [adaptado da referência 68].

Dessa forma, essa proteína representa um alvo extremamente interessante para se explorar a técnica de ligação cruzada como ferramenta experimental para determinação da estrutura terciária.

2. Objetivos

2.1 Objetivo Geral

Avaliar o uso dos dados experimentais obtidos através da técnica de ligação cruzada acoplada a espectrometria de massas em modelagem molecular para elucidação da estrutura terciária de proteínas.

2.2 Objetivos Específicos

Realizar os experimentos de ligação cruzada e obter o mapa de restrições de distância;

Utilizar essas restrições para a geração de modelos;

Avaliar os modelos gerados e elencar possíveis estruturas para a STC1.

3. Procedimento Experimental

3.1 Expressão e purificação da STC1-HT

A STC1-HT (STC1 ligada a uma causa de Histidina) foi produzida no LNBio (Laboratório Nacional de Biociências) sob supervisão do Dr. Daniel Maragno Trindade. Já se encontrava disponível suspensão com baculovírus recombinante.

Células High Five™ (Invitrogen) foram adaptadas a cultura em suspensão em meio livre de soro (Gibco) com adição de 20mM de L-Glutamina (Gibco) e 1x de PenStrep (Gibco). O estoque de cultura de células foi mantido em uma incubadora (ThermoForma) a 27ºC. Para a produção da STC1-HT, células High Five foram ampliadas da cultura estoque para uma densidade de células de aproximadamente $1x10^6$ em dois frascos Erlenmeyer de 2 L contendo 500 mL e incubados em shaker a 26ºC a 140 rpm. Após 20 horas de inoculação as células foram infectadas com o baculovírus recombinante.

Após 50 horas de infecção o meio de cultura foi centrifugado a 500 x g por cinco minutos e o sobrenadante contendo a STC1-HT secretada foi utilizado para purificação. Ao sobrenadante foi adicionado solução estoque de 1M de MES para uma concentração final de 50 mM e pH 6.5 (tampão IEX). A solução foi filtrada por uma membrana MCE de 0.45 μm (Fisherbrand) e a solução foi carregada em uma coluna previamente empacotada com SP Sepharose FF (Pharmacia Biotech/GE) em um fluxo de 1 mL/min utilizando uma bomba peristáltica (Bioogic LP – Biorad). A coluna foi transferia para um sistema de FPLC ÄKTA (GE) para a eluição da proteína utilizando um gradiente de 0-1 M de NaCl em tampão IEX. Frações eluidas a partir de 30 mS cm^{-1}, contendo a maior parte da STC1-HT foram agrupadas. Essa solução foi diretamente carregada em uma coluna pré-empacotada HisTrap crude FF 5 mL (GE) equilibrada com 50 mM MES pH 6.5, 500 mM NaCl (tampão de afinidade). Após injeção a amostra foi lavada com seis volumes da coluna (CV) com o tampão de afinidade, com três CV do tampão de afinidade contendo 250 mM de imidazol e finalmente com quatro CV de tampão de afinidade contendo 1 M de imidazol. Essas últimas frações que continham a STC1-HT foram agrupadas e concentradas para 500 μL utilizando um filtro de centrifugação Amicon Ultra-15 com membrana Ultracel-10 de 30000 NMWL (Millipore) em uma centrifuga refrigerada a 4 ºC e então aplicados em uma coluna Superdex 200 pg 16/60 (GE), pré-equilibrada com 60 mM MES, 600

mM NaCl pH 6,5 (tampão SizeEx) com um fluxo de 0,5 mL min^{-1}. A proteína foi eluída em um único pico entre 70 e 80 mL, concentrada e mantida em tampão SizeEx a 4 ^{o}C para posteriores análises.

3.2 Reações de ligação cruzada, alquilação e proteólise enzimática

Para a reação de ligação cruzada foram adicionados 300 uL de tampão fosfato 50 mM, pH = 7 à 100 uL de solução de STC1-HT 3 mg mL^{-1} (5 nmol) resultando em uma contração final de 12,5 μM. Uma alíquota de 37 μL de uma solução estoque de DSS 10 mg mL^{-1} (27 mM) preparado em DMF (N,N-dimetilformamida) anidra foi adicionada a solução de proteína de forma a se obter um excesso molar de 200 vezes entre o agente de ligação cruzada e a STC1-HT e mantendo a concentração de DMF menor que 10% (v/v). A reação foi incubada a 27 ^{o}C sob agitação de 300 rpm por duas horas. Após a reação de ligação cruzada foi adicionado 15 uL de uma solução estoque 10 mM de DTT (ditiotreitol) preparada em bicarbonato de amônio 100 mM (excesso molar de 30 x em relação a proteína), incubando-se em seguida por 30 minutos à 70 ^{o}C. A alquilação foi realizada adicionando-se 10 μL de solução estoque 50 mM de IAA (iodoacetamida) preparada em bicarbonato de amônio 50 mM, mantendo-se a solução por 30 min a temperatura ambiente e no escuro. Finalmente a proteólise enzimática foi realizada pela adição de 100 μL de solução de 100 mM de tampão bicarbonato em que estavam solubilizados 6 mg de tripsina porcina (Sequencing Grade Modified Trypsin, Promega), proporção de 1:50 (m/m) enzima:substrato, por 18h à 37 ^{o}C. O digesto tríptico foi dessalinizado através de um cartucho de extração em fase sólida (SPE) Oasis HLB que havia sido previamente equilibrado com 1 mL solução de ácido fórmico 0,1%, e o eluatos resultante foi concentrado até um volume de 50 μL e guardados refrigerados a 4 ^{O}C para posterior análise.

Para análise das ligações de dissulfeto foi realizada para a mesma quantidade de proteína somente a etapa de proteólise enzimática, como descrita acima.

3.3 Análises por LC-MS

Os peptídeos foram separados em um sistema de nano-LC (Easy-nLC 1000, Thermo Fisher Scientific) em uma coluna de fase reversa a um fluxo de 300 nL/min utilizando gradiente de 5 a 35% B (50 min); 35-70% B (2 min); e 70% B (8 min). Os solventes da fase móvel consistiam de A: 5% ACN, 0,1% ácido fórmico em água; e B: 95% ACN, 0,1% ácido fórmico em água. O sistema foi conectado em uma fonte de nano-ESI acoplada a um espectrômetro de massas Q-Exactive (Thermo Fisher Scientific. Os dados de espectrometria de massas foram adquiridos de maneira contínua durante todo o gradiente. Cada espectro de MS, adquirido no Orbitrap em uma faixa de massa m/z 400-1800 a uma resolução de 70000 (m/z 400), foi seguido por um modo de aquisição dependente de dados (DDA) controlado pelo *software* XCalibur 2.0 (Thermo Fisher Scientific). Os dez sinais mais intensos do espectro de massa foram selecionados para serem dissociados por colisão induzida (CID) e os fragmentos foram detectados pelo Orbitrap a uma resolução de 35000 (m/z 400).

A identificação da proteína e de modificações em peptídeos devido ao agente de ligação cruzada foi realizada empregando-se o sistema MASCOT v.2.2 (Matrix Science Ltd). Como parâmetros de busca foram selecionados digestão com tripsina permitindo-se um ou dois sítios de clivagem ignorados por peptídeo, carbamidometilação como modificação fixa e tolerância de massa de peptídeos e fragmentos de ±0,05 Da. Para busca de aminoácidos que sofreram modificação pelo agente de ligação cruzada hidrolizado, foram feitas buscas adicionando modificação variável de 156,0786 Da em resíduos de lisina, serina e treonina. As buscas foram realizadas utilizando conjuntamente o banco de dados do SwissProt e a sequência da STC1-HT adicionada em um banco de dados interno.

Os espectros candidatos peptídeos contendo ligação cruzada foram assinalados pelos *softwares Crux-for-Xlink*[69], *pLink*[70] e SIM[53] (*Spectrum Identification Machine,* programa desenvolvido em nosso laboratório), seguidos de validação manual.

3.4 Predições de modelos

Para a predição da modelos da STC1-HT foram utilizados três softwares: I-Tasser[71], Quark[72] e Rosetta[73–75]. Os dois primeiros foram utilizados em suas

plataformas online sendo necessária a sequência da STC1-HT e um arquivo com as restrições de distâncias.

Para a predição de modelos utilizando Rosetta foi utilizado o protocolo Ab Initio Relax encontrado nos pacotes de modelagem molecular Rosetta disponíveis a partir da versão 3.1. Os arquivos utilizados foram: 1) um arquivo FASTA contendo a sequência da STC1-HT; 2) um arquivo de predição de estrutura secundária utilizando PSIPRED[76] (http://bioinf.cs.ucl.ac.uk/psipred); 3) dois arquivos de fragmentos de 3-mer e 9-mer, respectivamente, adquirido através do servidor ROBETTA[77] (http://robetta.bakerlab.org); 4) um arquivo contendo as restrições de distância experimentais que no seguinte formato:

AtomPair {nome do átomo 1} {número do resíduo 1, cadeia ID1} {nome do átomo 2} {número do resíduo 2, cadeia ID2} FLAT_HARMONIC $\{x_o\}$ {desvio padrão} {tolerância}

Em que x_o = 15,0, tolerância = 15,0 e σ= 1,0. Essa função garante que modelos sejam penalizados em seus scores quando a distância Euclidiana entre dois átomos descritos excede 30,0 Å.

Neste trabalho foram gerados quatro conjuntos de modelos para a STC1-HT utilizando o Rosetta, aqui chamados de **A, B, C** e **D**:

A: modelos utilizando todos os arquivos descritos – conjunto de dados no arquivo de restrições parcial;

B: mesma condição de **A** sem utilização do arquivo (2);

C: mesma condição de **B** com arquivo (1) contendo a sequência da proteína selvagem (sem causa de histidina) da STC1 e com arquivo de restrições completo;

D: mesma condição de **C**, sem utilizar o protocolo *relax* do Rosetta.

3.5 Critérios de seleção de modelos

Cada conjunto de modelos foi avaliado de acordo com: 1) a distância euclidiana entre os Cα envolvidos nas ligações de dissulfeto; 2) a distância euclidiana entre os

Cα envolvidos nas ligações cruzadas experimentais; 3) a distância topológica acessível ao solvente. Esses dados foram gerados utilizando o programa Xwalk[78].

3.6 Agrupamentos de modelos semelhantes

Dentre os modelos que respeitaram as restrições experimentais, foram selecionados os 500 modelos de menor *score* do Rosetta de cada conjunto. Esses modelos foram agrupados utilizando a ferramenta MaxCluster, e foram testados três índices de avaliação de similaridade de estruturas de forma que pelo menos metade das estruturas pertencessem a um grupo: a) GDT *score* (Tm = 80); b) MaxSub *score* (Tm = 0,65) e c) RMSD (Tm = 11,5 Å). Foi utilizado o algoritmo do vizinho mais próximo restrito (Restricted *Nearest Neighbour clustering*) para o agrupamento.

3.7 Simulações por Dinâmica Molecular e Análise de trajetória

Simulações de dinâmica molecular (MD) foram realizadas para STC1-HT a fim de se avaliar o comportamento dessas estruturas em solução. Todas as simulações foram realizadas com NAMD[79]. VMD[80] foi utilizado para visualização e para a preparação das figuras. Os parâmetros CHARMM27[9] foi utilizado para a proteína. O modelo TIP3P foi utilizado para a água[81]. As estruturas foram solvatadas com Packmol[82,83] com água, íons sódio e cloreto para neutralizar a caixa de simulação. Os sistemas completos possuíam entre 60 e 70 mil átomos. As estruturas solvatadas foram submetidas aos seguintes passos de minimização e equilibração: 1) 40 ps de MD mantendo todos os átomos da proteína fixa; 2) 40 ps de MD mantendo todos os átomos da cadeia principal da proteína fixa; 3) 40 ps de MD sem manter nenhum átomo fixo. Foram realizadas simulações de 25 ns a partir da condição final do passo 3). As simulações foram realizadas como passo de 2 fs, a temperatura (298 K) e pressão (1 bar) constante. Um limite de 12 Å para as interações de vdW foi utilizado. Interações de longo alcance de potencial eletrostático foram calculados utilizando o método de soma de Ewald. A trajetória foi gravada a cada 1000 passos. Todas as simulações foram realizadas em um computador Xeon® X5650 2,67 GHz, 24 Gb de memória RAM e placa GPU NVDIA® C2075.

A análise da trajetória foi realizada com o MDLovoFit. A construção dos gráficos foi realizada com MatLab R2014a[84].

4. Resultados e Discussão

4.1 Obtenção de dados de restrição para a STC1-HT

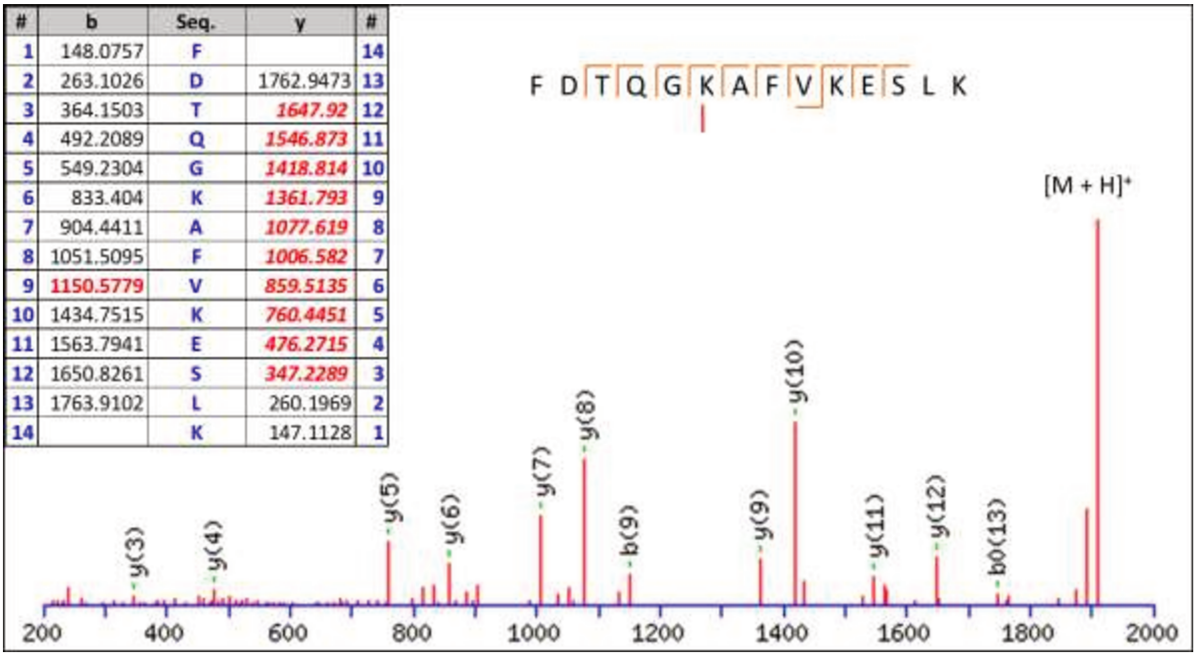
A construção da STC1-HT possui 260 resíduos de aminoácidos (**Figura 12**), em que os 230 primeiros dizem respeito a STC1 selvagem sem os 17 primeiros resíduos, que correspondem ao peptídeo sinal e não está presente na STC1-HT que é secretada para o meio extracelular durante a expressão. Os outros 30 resíduos de aminoácidos presentes no C-terminal correspondem a porção His-Tag incorporada a proteína para a etapa de purificação por afinidade.

A fim de se obter um número de restrições experimentais razoáveis, utilizando-se DSS como ALC, verificou-se em um primeiro momento o número de lisinas que essa proteína possuía, uma vez que esse ALC reage preferencialmente com aminas primárias e, assim, com as cadeias laterais de lisina e com o N-terminal e, em um segundo momento, a presença de treoninas e serinas, que possuem grupos hidroxila que podem reagir dependendo do microambiente químico a que estão expostas na estrutura terciária da proteína. A presença de 13 lisinas e 37 resíduos de serina ou treonina torna a STC1-HT uma proteína passível de ser avaliada frente a reação com o DSS. De fato, os seis resíduos reativos que estão presentes na porção C-terminal e que não estão presentes na STC1 selvagem não possui interesse do ponto de vista da elucidação estrutural da STC1.

A digestão enzimática com tripsina garante especificidade de clivagem após resíduos de arginina e lisina, desde que elas não sejam seguidas por prolina. A clivagem também não ocorrerá se o resíduo de lisina estiver modificado pelo ALC, pois a enzima não o reconhece mais o sítio, que deixou de ser básico, da cadeia lateral desse resíduo.

Figura 12: Sequência dos resíduos de aminoácidos que compõe a STC1-HT. Em vermelho estão destacados os 17 resíduos de lisina, que reagem preferencialmente com o DSS; em verde estão destacados os 38 resíduos de treonina e serina, que são sítios que também podem reagir com o mesmo reagente.

A partir análise das corridas de LC-MS e LC-MS/MS com auxílio do programa MASCOT, é possível identificar resíduos de lisina, serina ou treonina que foram modificados por uma porção do DSS mas o qual sofreu hidrólise na outra porção. O que ocorre é que a modificação causada funciona exatamente como uma modificação pós-traducional convencional, sendo necessária somente a análise convencional dos peptídeos adicionando-se como modificação variável a massa de 156,0786 Da. Como pode ser observado em um espectro desse tipo na **Figura 13**, a fragmentação convencional em íons **–b** e **–y** permite identificar os dois sítios de lisina nos quais ocorreram a modificação com o DSS hidrolisado.



#	b	Seq.	y	#
1	148.0757	F		14
2	263.1026	D	1762.9473	13
3	364.1503	T	*1647.92*	12
4	492.2089	Q	*1546.873*	11
5	549.2304	G	*1418.814*	10
6	833.404	K	*1361.793*	9
7	904.4411	A	*1077.619*	8
8	1051.5095	F	*1006.582*	7
9	1150.5779	V	*859.5135*	6
10	1434.7515	K	*760.4451*	5
11	1563.7941	E	*476.2715*	4
12	1650.8261	S	*347.2289*	3
13	1763.9102	L	260.1969	2
14		K	147.1128	1

Figura 13: Espectro de ESI-MS/MS de um peptídeo modificados pelo DSS em sua forma hidrolisada obtido pela ferramenta MASCOT durante a busca de identificação da SYC1-HT. Os íons fragmentos estão anotados conforme a série –b e –y. O traço vermelho na sequencia indica a lisina na qual ocorreu a modificação com ALC.

Os programas atualmente disponíveis para a identificação de espécies inter- e intra-peptídeos como Crux e plink, fornecem uma lista de candidatos através de um *score* próprio do programa. Entretanto, a análise de cada espectro de MS/MS deve ser feita manualmente e separadamente para não incorrer em atribuições erradas dessas espécies de interesse. O programa SIM é uma ferramenta atrativa nesse sentido por permitir uma avaliação não somente numérica do *score* atribuído a um candidato mas também por permitir que o usuário faça uma avaliação visual do espectro atribuído, verificando a presença de sinais intensos não atribuídos a nenhum fragmento ou a baixa cobertura de fragmentação, sinais de que o espectro não corresponde àquela atribuição ou, ao contrário, que todos os sinais intensos estão associados a um fragmento e que a cobertura é suficiente para validar a identificação. Dessa forma, análises que poderiam demorar semanas podem ser realizadas em alguns dias[53].

A **Figura 14A** mostra um espectro de fragmentação de uma espécie inter-peptídica envolvendo dois sítios de lisina e a **Figura 14B** mostra um espectro de fragmentação do mesmo tipo mas envolvendo a cadeia lateral de um resíduo de

serina e o N-terminal. É possível observar que ambos os espectros possuem sinais de fragmentação que delimitam toda a extensão de ambas as cadeias polipeptídicas e que todos os sinais mais intensos do espectro estão atribuídos. Também é possível observar para o espectro **a** a presença dos sinais 222,149 e 305,223, correspondentes a íons marcadores envolvendo resíduos de lisina ligados ao DSS.

A **Figura 15A** destaca a identificação de uma espécie simétrica, correspondente a uma região na porção C-terminal onde há uma interação física da espécie homodimérica. Esta informação juntamente com a ligação de dissulfeto anteriormente caracterizada (cisteína 185) compõe um componente-chave para a compreensão da topologia do homodímero. De maneira semelhante, foi possível mapear algumas ligações de dissulfeto já publicadas (**Figura 15B**). Elas funcionam como ligações cruzadas naturais e também apresentam fragmentação convencional.

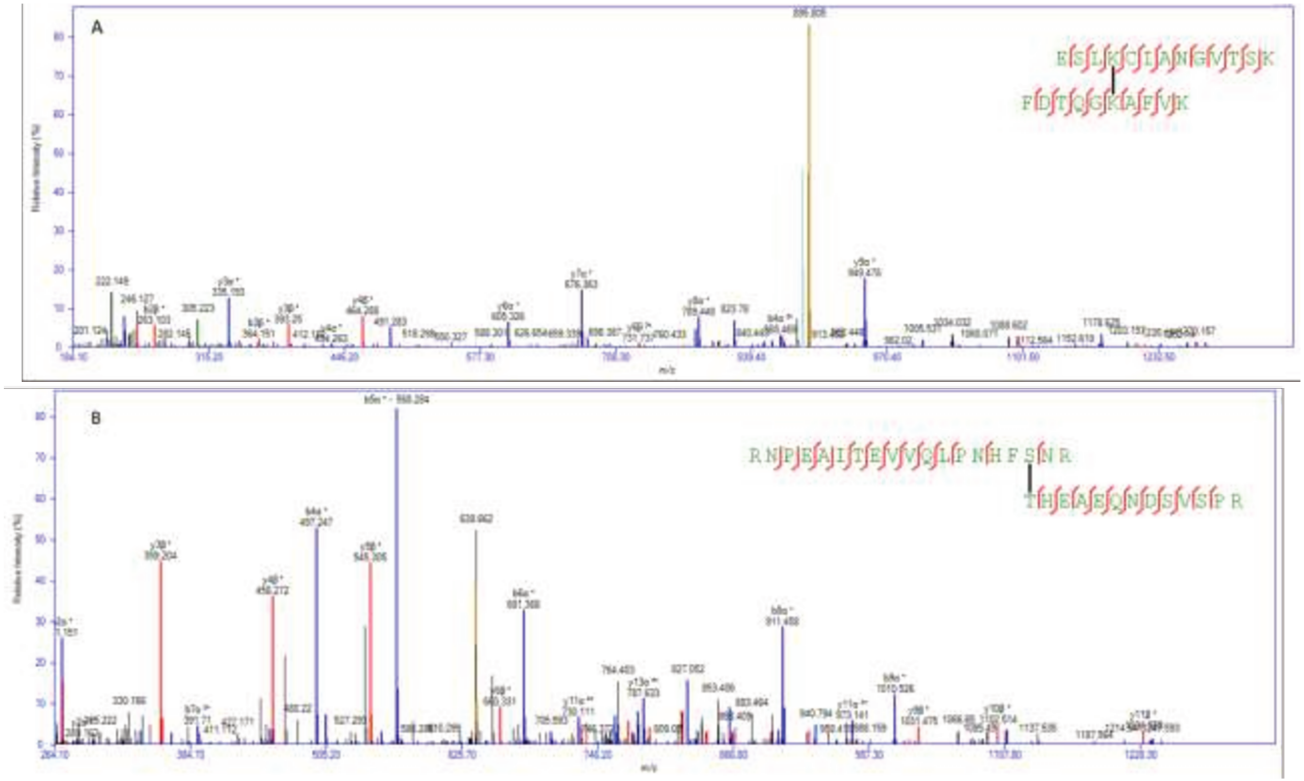


Figura 14: Espectros de ESI-MS/MS para espécies inter-peptídicas envolvendo em (A) dois resíduos de lisina, encontrados convencionalmente em experimentos de ligação cruzada e, em (B), uma espécie envolvendo a treonina, que também é o N-terminal da STC1-HT e um resíduo de serina, menos comum nesse tipo de experimento. Para o espectro em (a) é possível identificar os íons marcadores, destacados pelos sinais em verde. Os íons fragmentos estão anotados conforme a série –b e –y.

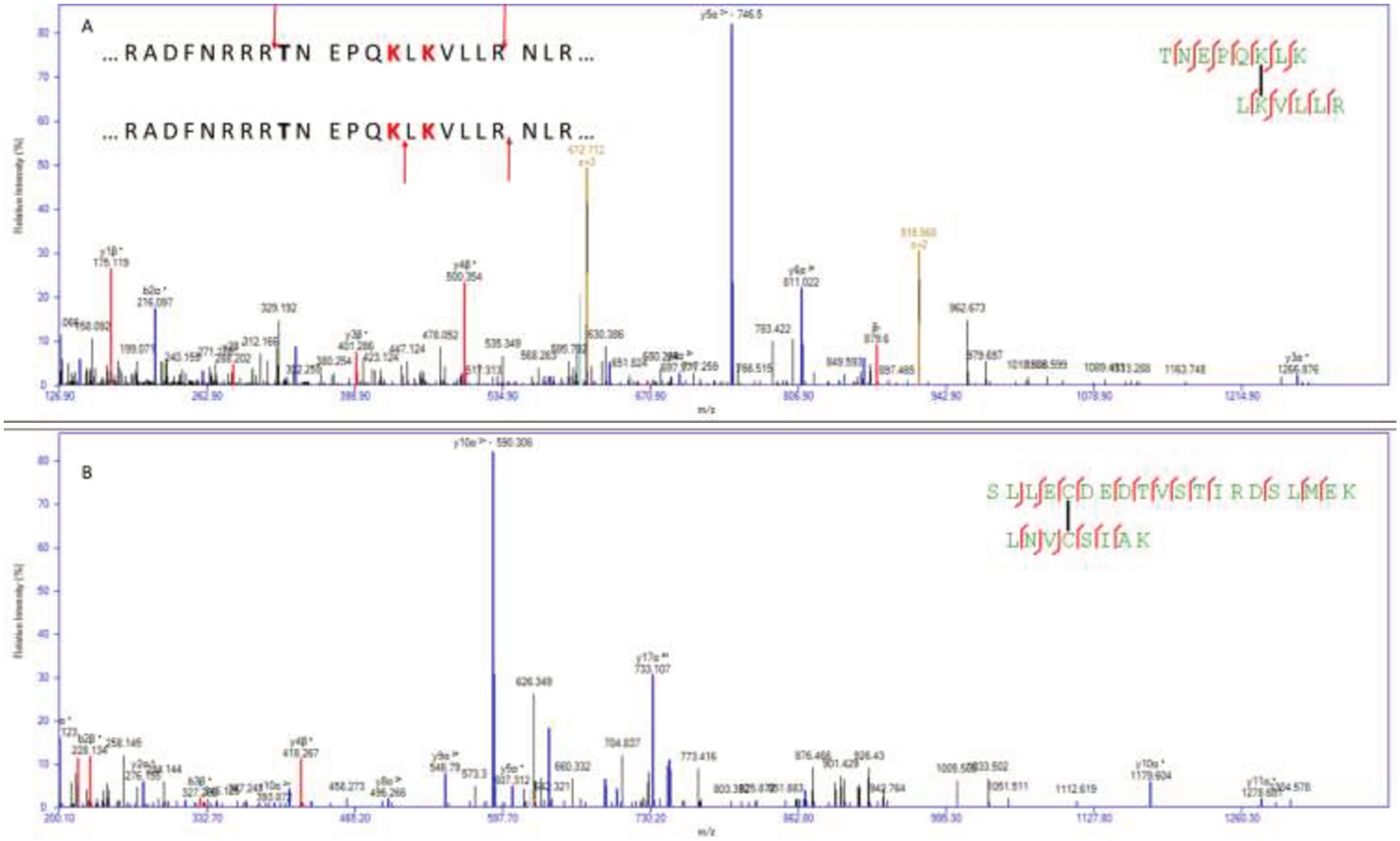


Figura 15: Espectros de ESI-MS/MS para espécies inter-peptídicas em (A) envolvendo duas lisina de cadeias distintas do homodímero, informação importante para se estabelecer a topologia de interação entre os monômeros; É mostrado no mesmo espectro na parte superior à esquerda um esquema das porções da sequência com as lisinas envolvidas na ligação cruzada em vermelho e as setas azuis indicando os sítios de clivagem por proteólise com tripsina; em (A) envolvendo duas cisteínas, uma das cinco ligações cruzadas naturais presentes na STC1.

Após a análise de todos os espectros de fragmentação candidatos a espécies inter- e intra-peptídicas, pode-se montar um mapa de restrições de distância, mostrados na **Figura 16**. No total foram possíveis identificar além das cinco ligações de dissulfeto previamente caracterizadas, e a ligação de dissulfeto responsável pelo homodímero, 31 restrições de distância envolvendo as cadeias laterais de resíduos de lisina, serina e treonina com a cadeia espaçadora do DSS, 11 restrições envolvendo o N-terminal, uma restrição relacionada a espécie inter-cadeias na porção C-terminal e 24 resíduos modificados com a espécie *dead end*. Como esperado todos os resíduos que estão envolvidos em espécies de ligação cruzada possuem um correspondente com a espécie hidrolisada.

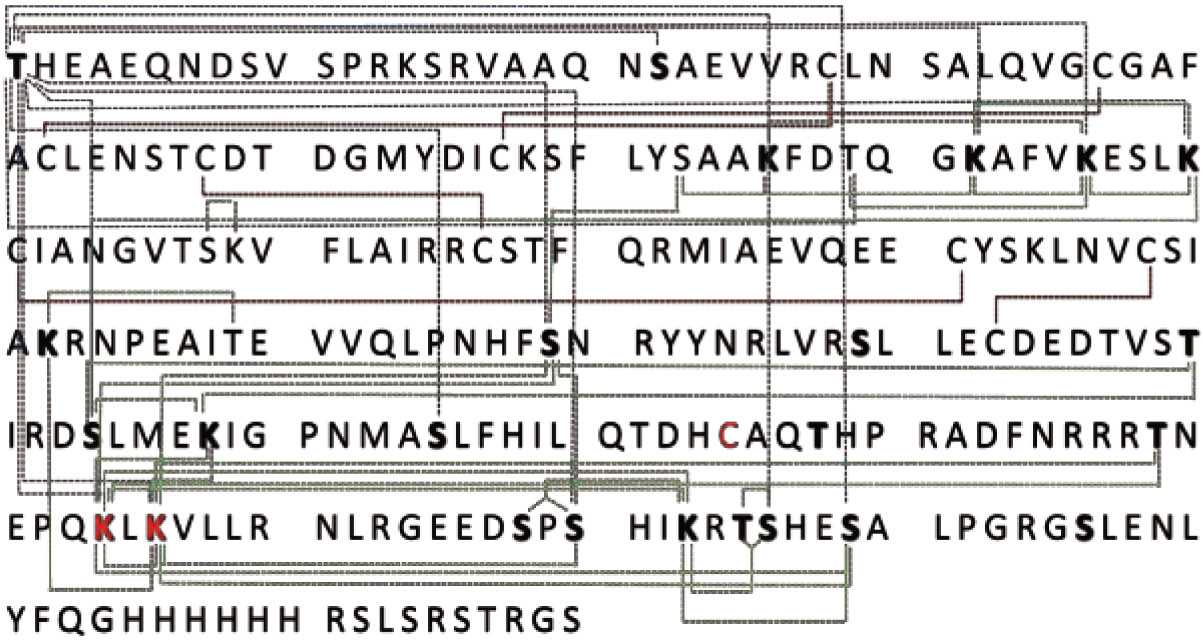


Figura 16: Mapa de ligações cruzadas para a STC1-HT. As cinco ligações de dissulfeto estão mostradas nas linhas tracejadas em vermelho; as ligações cruzadas envolvendo resíduos diferentes do N-terminal são mostradas nas linhas tracejadas em verde; as ligações cruzadas envolvendo o N-terminal estão mostradas nas linhas tracejadas em azul. Em negrito são destacados resíduos em que se encontraram modificados pelo DSS hidrolisado e em vermelho estão destacados os resíduos envolvidos na espécie dimérica.

4.2 Avaliações dos modelos frente as restrições experimentais

A fim de avaliar a qualidade dos modelos que foram gerados (discutidos mais a frente), utilizou-se a distância euclidiana entre os C_α dos resíduos envolvidos na ligação cruzada, calculadas utilizando o programa Xwalk. A STC1 possui cinco ligações de dissulfeto (ligações cruzadas naturais). Verifica-se nas estruturas que possuem ligações de dissulfeto depositadas no PDB[85], que as distância entre os C_α das cisteínas envolvidas variam entre 3,8 e 6,8 Å (**Figura 17**), sendo que aqui estabeleceu-se o limite superior de 7 Å para a avaliação com relação a estas restrições.

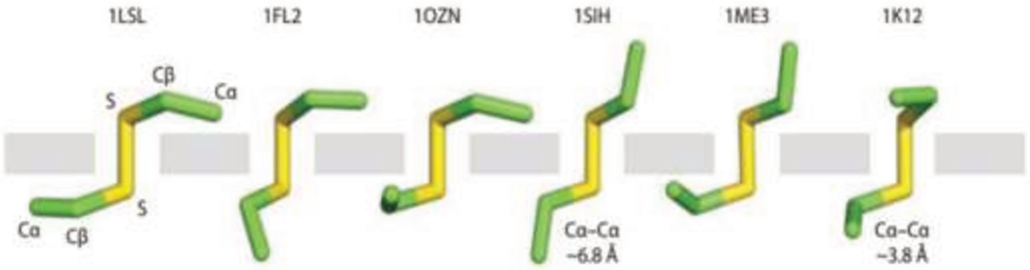


Figura 17: Representação de ligações de dissulfeto extraídas de estruturas de proteínas. Os códigos das proteínas do PDB estão indicados acima de cada representação. As diferentes conformações indicam que as distâncias Cα-Cα estão na faixa entre 3,8 Å até 6,8 Å.[85]

Ao contrário das ligações de dissulfeto, que ocorrem naturalmente, a distância entre lisinas envolvidas em uma ligação cruzada não pode ser avaliada diretamente das estruturas depositadas no PDB. Tradicionalmente na literatura, utiliza-se uma distância máxima entre os N_ζ de 11,4 Å e entre C_α de 24,4 Å para resíduos de lisina ligados pelo ALC. Entretanto, nunca havia sido conduzido um estudo sistemático a esse respeito e, muitas vezes, observava-se experimentalmente que alguns resíduos envolvidos em ligação cruzada, ou seja, que estavam espacialmente próximos, possuíam estruturas cristalográficas em que esses mesmos resíduos estavam distantes além do limite de espaçamento desse ALC.

Recentemente foi publicada[86] uma avaliação de um banco de dados de simulações de dinâmica molecular de diversas estruturas, representativas dos enovelamentos disponíveis no PDB, frente a dados coletados na literatura com relação a ligação cruzada realizadas com DSS. A **Figura 18** ilustra como, devido a

flexibilidade da cadeia lateral, a distância entre os C_α é uma medida mais precisa no sentido de contemplar a maior parte das distâncias avaliadas em uma determinada estrutura, de forma estática. Enquanto que o valor máximo de 11,4 Å entre N_ζ é contemplada por apenas 20% das distâncias avaliadas na estrutura, o limite superior de 24,4 Å entre C_α é respeitada por mais de 80% das distâncias avaliadas na estrutura. Os autores recomendam, por fim, o limite superior de 30 Å entre C_α como uma distância apropriada para a análise de um modelo envolvendo ligação cruzada com DSS. Vale lembrar que os dados avaliados pelos autores se referem à distância euclidiana, sendo que a distância topológica pode ser muito discrepante quanto a esse valor.

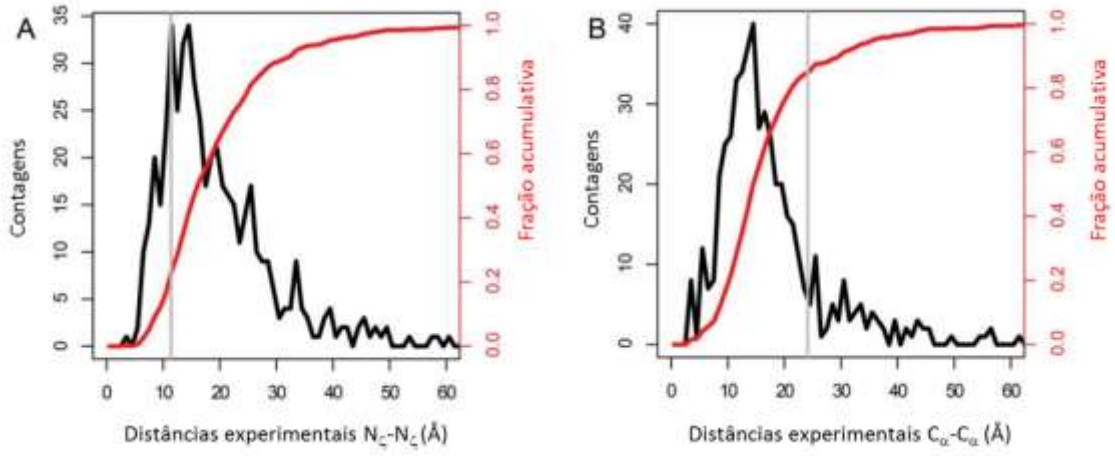


Figura 18: Distâncias entre átomos envolvidos em experimentos de ligação cruzada e proteínas com estruturas experimentais conhecidas. As linhas em preto mostram as distribuições, as linhas em vermelho as distribuições cumulativas e a linha em cinza indica a distância convencional atribuída para esse ALC de 11,4 Å para distâncias entre N_ζ e de 24,4 Å para distâncias entre C_α. O número total de restrições incluído nesse conjunto de dados foi de 502. (A) Distâncias experimentais entre N_ζ– N_ζ . A tendência central é consideravelmente maior que 11,4 Å. (B) Distâncias experimentais entre Cα-Cα. A distribuição de distância tem seu máximo abaixo do valor teórico para essa distância. [adaptado da referência 86]

Os desvios que podem estar presentes em um modelo frente aos dados de restrições de distância experimentais podem ser atribuídos à natureza intrinsecamente dinâmica de uma proteína em solução, sendo que o dado obtido

reflete as várias conformações possíveis. Assim, é aceitável que um mesmo resíduo possa estar envolvido em duas restrições de distância que são mutuamente exclusivas em uma única estrutura proposta pois elas podem ter sido formadas em solução em duas moléculas de proteínas com conformações distintas. É de se esperar que resíduos envolvidos em ligações cruzadas presentes em regiões mais rígidas da proteína apresentem nenhum ou poucos desvios quanto a distância limite esperada enquanto que regiões mais flexíveis e pouco estruturadas possam apresentar valores muito maiores.

De posse desse conjunto de restrições experimentais é possível avaliar como os programas de modelagem se comportam frente a esses dados. I-Tasser e Quark foram escolhidos em um primeiro momento pois, além do já citado destaque nas avaliações promovidas pelo CASP, oferecem plataformas online com a possibilidade de utilizar como arquivo de entrada, juntamente com a sequência da proteína de interesse, uma lista de restrições entre átomos da proteína com distâncias euclidianas que os correlacionam. Ainda vale ressaltar que o Quark oferece a possiblidade de se introduzir somente 200 resíduos de aminoácidos o que, no caso da STC1-HT, promoveria um modelo incompleto. I-Tasser gera como resultado final cinco modelos melhores avaliados, enquanto que o Quark fornece dez modelos em seu relatório final. Os resultados desses 15 modelos quando avaliados quanto a distância euclidiana relacionadas as cinco ligações de dissulfeto são mostrados na **Figura 19**.

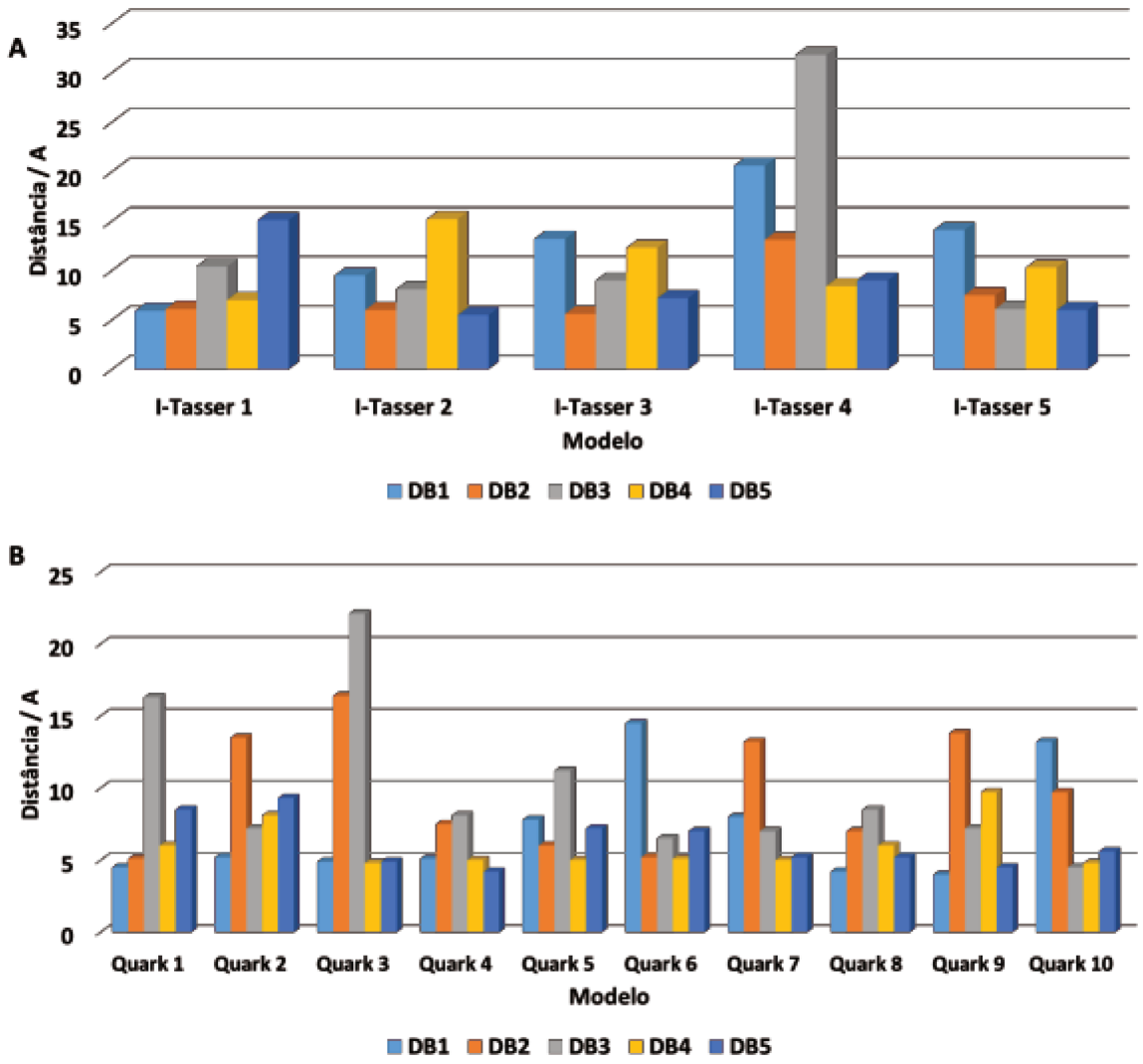


Figura 19: Valores encontrados para as distâncias entre os Cα das ligações de dissulfeto para os cinco modelos gerados pelo I-Tasser (A) e para os dez modelos gerados pelo Quark (B). Grandes desvios em relação a distância esperada foram encontrados para ambos os programas.

Ambos os programas geraram modelos que possuíam uma ou mais distâncias euclidianas relacionadas as ligações de dissulfeto acima do limite esperado. Outras abordagens foram utilizadas na plataforma online como, por exemplo, reduzir o número de aminoácidos ou diminuir o número de restrições. Nesses casos, mas resultados semelhantes foram obtidos pois, provavelmente, o algoritmo possui um critério que atribui peso pequeno para os dados de restrições e, devido à pouca flexibilidade na interface dos programas para o usuário não foi possível obter dados relevantes utilizando-os.

No final e 2013, foi publicado um artigo que utiliza o protocolo *ab initio* do pacote Rosetta associado com dados de restrições de distância[87]. Neste trabalho,

além das informações de restrições de distância, o sistema de interesse de se obter um modelo (proteína IgBP1 humana) possuía um domínio N-terminal homólogo de rato com estrutura disponível no PDB. Como mostrado posteriormente, na análise de agrupamento realizada em nosso trabalho, o RMSD não é o melhor parâmetro disponível para comparação estrutural, apesar de rápida, pode incorrer a associação de estruturas que não são de fato similares.

O pacote Rosetta oferece a vantagem de poder ser executado localmente e seus parâmetros manipulados pelo o usuário. Cada passo de geração de modelos acompanhou a obtenção das restrições de distância, sendo que cada grupo de modelos gerados possuía um arquivo de restrições mais completo que o anterior.

Foram gerados nas condições **A** e **B,** 35 mil modelos, na condição **C**, 15 mil modelos, e na condição **D**, 21 mil modelos, totalizando mais de 105 mil modelos. A **Figura 20** mostra a avaliação do número de modelos que respeitaram as cinco restrições relacionadas as ligações de dissulfeto e também das 31 restrições encontradas para as espécies de ligação cruzada. Como os números de cada conjunto são distintos, os valores foram normalizados para facilitar a comparação.

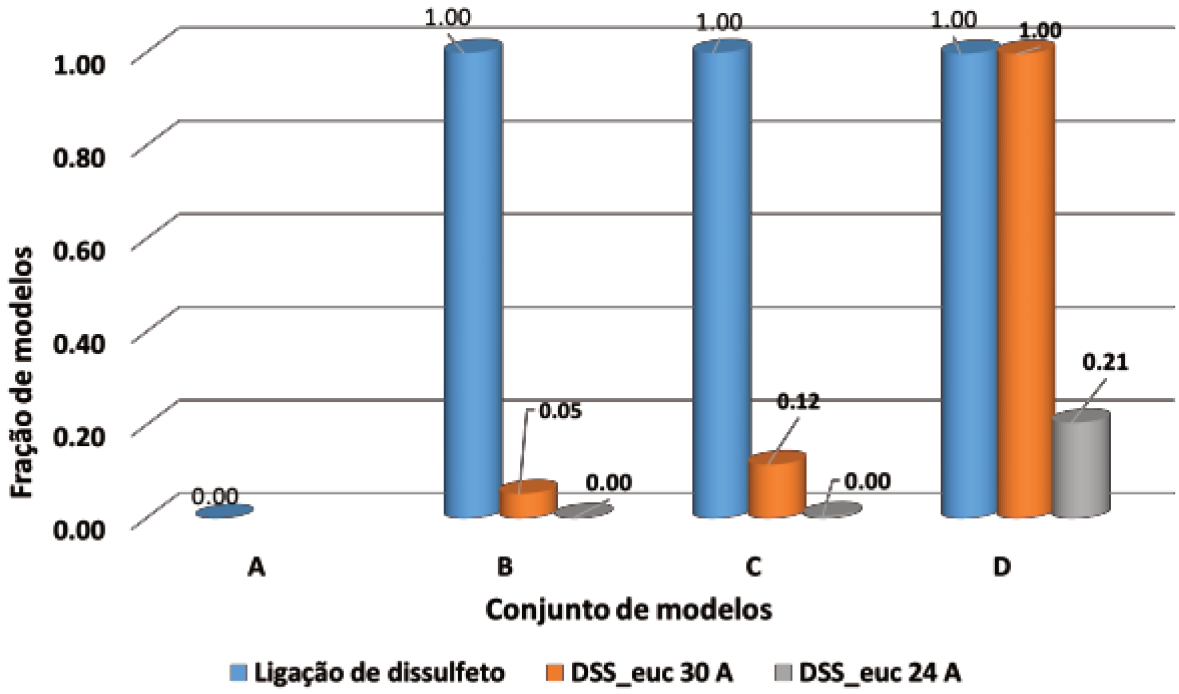


Figura 20: Gráfico mostrando a fração de modelos que obedecem as restrições de distância relacionadas com as ligações de dissulfeto (em azul) e pelas restrições experimentais obtidas utilizando-se DSS com diferentes limites superiores para cada conjunto gerado pelo Rosetta. As ligações de dissulfeto são respeitadas por praticamente todos os modelos entre os conjuntos B, C e D, enquanto que as restrições do ALC são eficientes na seleção de modelos nos conjuntos B e C.

A primeira observação é que para o conjunto **A** nenhum modelo respeitou as cinco restrições relacionadas às ligações de dissulfeto, o que levou a explorar outras opções do programa. Para o conjunto **B**, decidiu-se retirar o arquivo de predição de estrutura secundária e, neste caso, praticamente todos os modelos passaram a ser gerados de forma a respeitar essas cinco restrições. Além disso, para o conjunto **B** ainda verificou-se que pouco mais de 5% das estruturas respeitam todas as restrições de distâncias relacionadas com o DSS dentro do limite superior de 30 Å e que apenas aproximadamente 0,2% representam estruturas em que as distâncias estavam limitadas a 24 Å. Esses dois conjuntos foram em dois meses e meio cada. O conjunto **C** contém um conjunto de modelos gerados com 30 resíduos de aminoácidos a menos (correspondentes à cauda de histidina), o que diminuiu em torno de 20% o tempo computacional para geração de modelos, e um arquivo de restrições de distância que possuía todas as 31 restrições não pertencentes à região N-terminal que foram obtidas experimentalmente. O que se observa é um aumento

no número de modelos que respeitam as restrições de distância (aproximadamente o dobro). Não é possível determinar se essa diferença de aproximadamente 100% em relação ao conjunto **B** é um efeito do arquivo de restrições ser completo ou da diminuição do número de resíduos de aminoácidos mas, provavelmente, há um efeito conjunto de ambas as variáveis. O conjunto **D** é talvez o mais intrigante pois todos os modelos respeitaram todas as restrições dentro do limite esperado de 30 Å e mais de 20% estavam com as restrições abaixo de 24 Å. Como esses modelos são gerados utilizando os átomos da cadeia principal e substituindo os átomos da cadeia lateral por um centroide com raio de Van der Walls que procura simular a densidade eletrônica da cadeia lateral, esses modelos são considerados de baixa resolução e são gerados em torno de 70% mais rápidos que os modelos que são submetidos à etapa de refinamento pelo programa. Provavelmente as restrições de distância são fielmente reproduzidas nessa etapa para então terem papel secundário na etapa de refinamento.

Os modelos dos conjuntos **B** e **C** foram selecionados para se avaliar a influência do número de restrições na seleção de modelos. A **Figura 21** mostra o comportamento para esses dois conjuntos frente à seleção com limite superior de 30 Å entre C_α. O conjunto **B** mostra a seleção de estruturas utilizando até 24 restrições de distância é pouco discriminatória, mas que ela passa a ser exponencialmente eficiente conforme aumentam o número de restrições. Este comportamento é ainda mais drástico para o conjunto **C**, onde a eficiência na seleção aumenta de forma ainda mais drástica com o aumento do número de restrições a partir de 27 restrições.

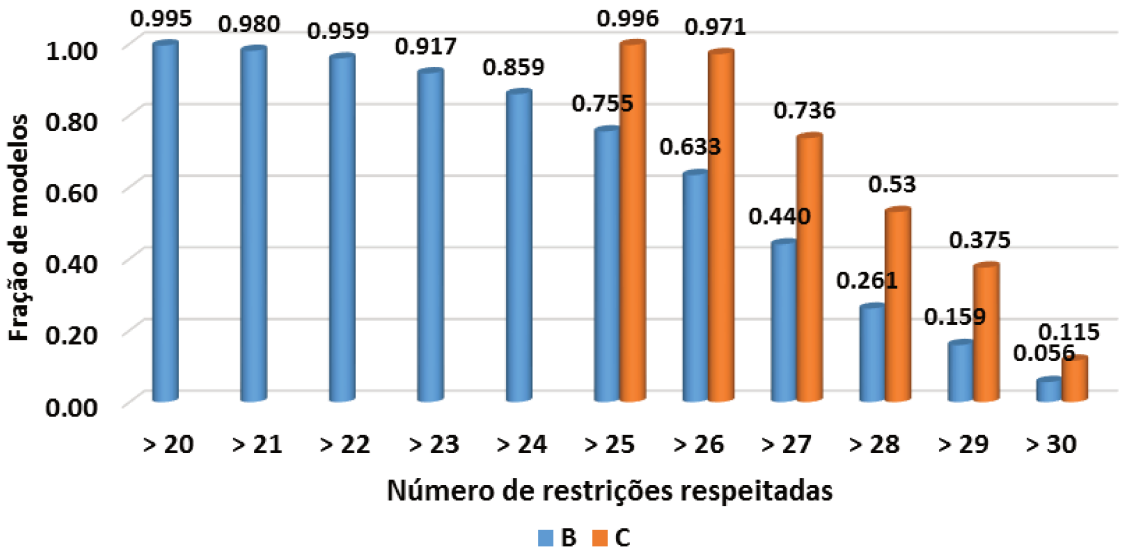


Figura 21: Influência do número de restrições experimentais utilizadas na seleção de modelos para os conjuntos **B** e C. É claramente observado que o maior número de restrições possui efeito sobre a seleção de modelos chegando a possuir um comportamento exponencial em determinados pontos (a partir de 25 restrições para o conjunto **B** e 27 restrições para o conjunto C).

Para saber se existem restrições que sejam mais importantes que outras na seleção de modelos, os mesmos dois conjuntos foram analisados para cada restrição individualmente. Os resultados apresentados na **Figura 22** mostram que as restrições de distância são divididas em dois grupos: 1) restrições de curta distância, normalmente encontradas em espécies intra-peptídicas, e que envolvem resíduos de aminoácidos distantes em até dez posições na sequência da proteína; e 2) restrições de longa distância, que são atribuídas a espécies inter-peptídicas, e que envolvem resíduos de aminoácidos distantes mais do que dez posições na sequência da proteína. Na **Figura 22**, as restrições estão em ordem crescente de separação dos aminoácidos na sequência da proteína. É interessante notar que as restrições de curta distância são obedecidas por todos os modelos enquanto que, de fato, as restrições mais importantes para a seleção de modelos são as restrições de longa distância (> 10 resíduos). Além disso, as restrições 27, 28 e 29 são as mais seletivas. Todas essas três restrições envolvem a serina-139 com três resíduos próximos da porção C-terminal, o que mostra que restrições mutuamente exclusivas são mais difíceis de serem modeladas, por provavelmente se tratar de regiões mais flexíveis. Essas restrições devem portanto, ser geradas a partir de diferentes confôrmeros da

proteína e essa possibilidade pode ser avaliada por meio de simulações de dinâmica molecular.

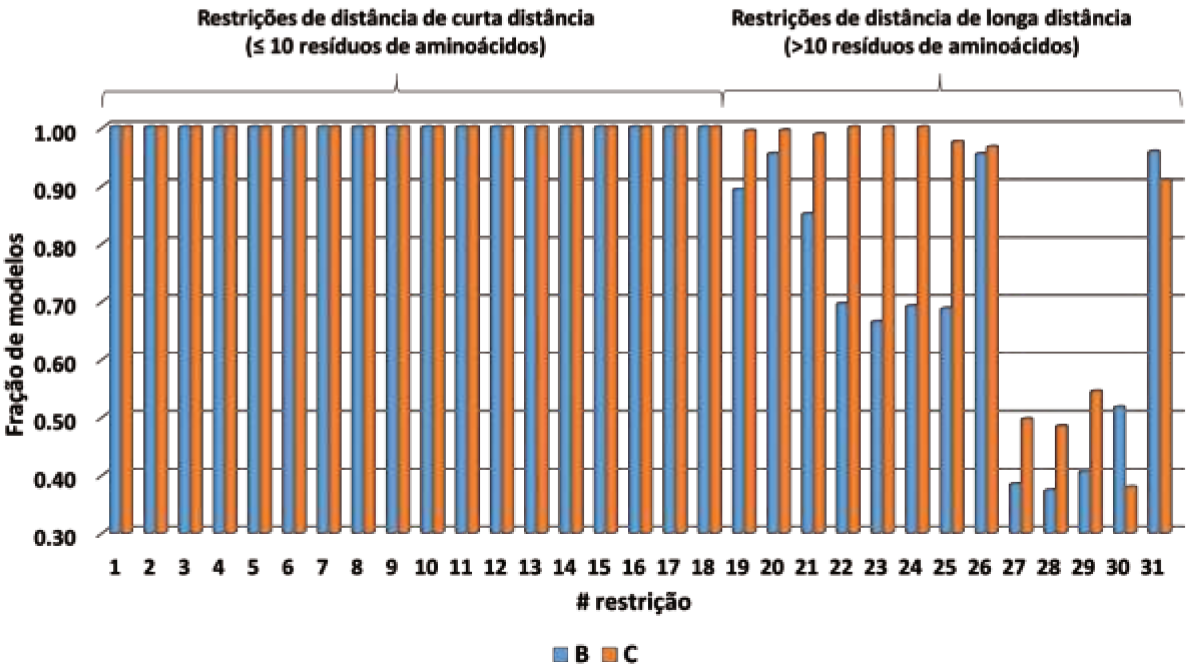


Figura 22: Relação entre o número de modelos filtrados por restrição. Restrições de curta distância (≤ 10 resíduos de aminoácidos) são obedecidas por todos os modelos, enquanto que restrições de longa distância são as informações mais importantes para a seleção de modelos.

4.4 Simulações de Dinâmica Molecular

As simulações de dinâmica molecular fornecem um meio de avaliar o comportamento de estruturas em solução em função do tempo. Alguns modelos estruturalmente distintos foram submetidos a dinâmicas de 25 ns e avaliados quanto à flutuação quadrática média (RMSF) para os 20 ns finais de simulação. O RMSF mede as flutuações locais de uma proteína por resíduo e pode ser calculado de forma semelhante ao RMSD como:

$$RMSF = \left(\frac{1}{T}\sum_{t_j}^{T}\left(x_i(t_j) - x_o\right)^2\right)^{1/2}$$

Onde:

T é o tempo durante o qual se faz a média e x_0 é a posição de referência, normalmente a posição média da partícula durante o tempo T.

A **Figura 23** destaca os resultados de RMSF para as cinco estruturas simuladas. Todas as estruturas apresentam grande flexibilidade na porção referente ao His-Tag, localizado na região C-terminal, o que é esperado, uma vez que se trata de uma porção não relacionada com a STC1 selvagem. Também é possível observar para as estruturas de A-C que a maior flexibilidade está contida na porção C-terminal desses modelos e que a porção N-terminal é relativamente rígida. Esses dados corroboram a maior os dados discutidos no final da sessão anterior, em que a serina-139 está relacionada com três espécies de ligação cruzada de longa distância com resíduos distintos da porção C-terminal: lisinas-204 e 206, e serina 220. Entretanto, contrariam o que foi encontrado nos experimentos de ligação cruzada em relação às 11 restrições envolvendo o N-terminal com resíduos em diversos pontos da proteína.

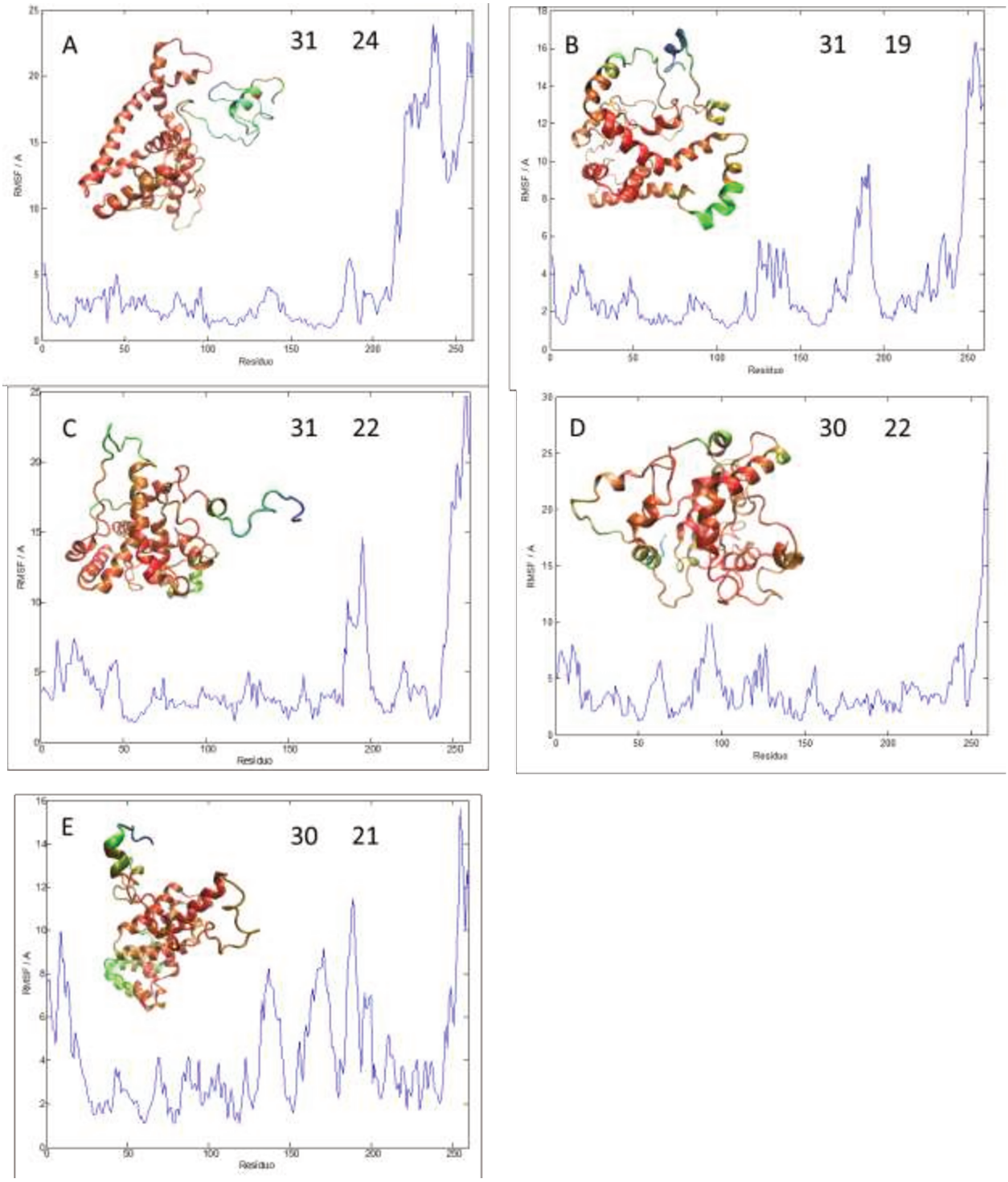


Figura 23: Gráficos representando os resultados de RMSF para cinco estruturas da STC1-HT. As estruturas estão coloridas de forma gradiente em que regiões mais rígidas são representadas na cor vermelha e regiões mais flexíveis são representadas na cor azul. Os valores destacados nos gráficos representam, respectivamente, o número de restrições euclidianas e topológicas obedecidas pelo modelo.

4.4 Distâncias euclidiana e topológica

A predição estrutural de proteínas utilizando-se dados de restrição de distância é realizada atribuindo-se sempre uma distância limite de natureza euclidiana entre dois pares de resíduos de aminoácido. O programa Xwalk é uma ferramenta que permite avaliar em larga escala esses modelos nesse sentido. Entretanto, a distância euclidiana não representa a distância seguida pelo ALC, uma vez que o ALC tem que se aproximar dos resíduos pela superfície da proteína e não através dela. Uma medida não linear que representa o caminho acessível entre os dois resíduos na superfície de uma proteína é, assim, uma distância, chamada de distância topológica. Esse valor também pode ser obtido através do programa Xwalk (**Figura 24**).

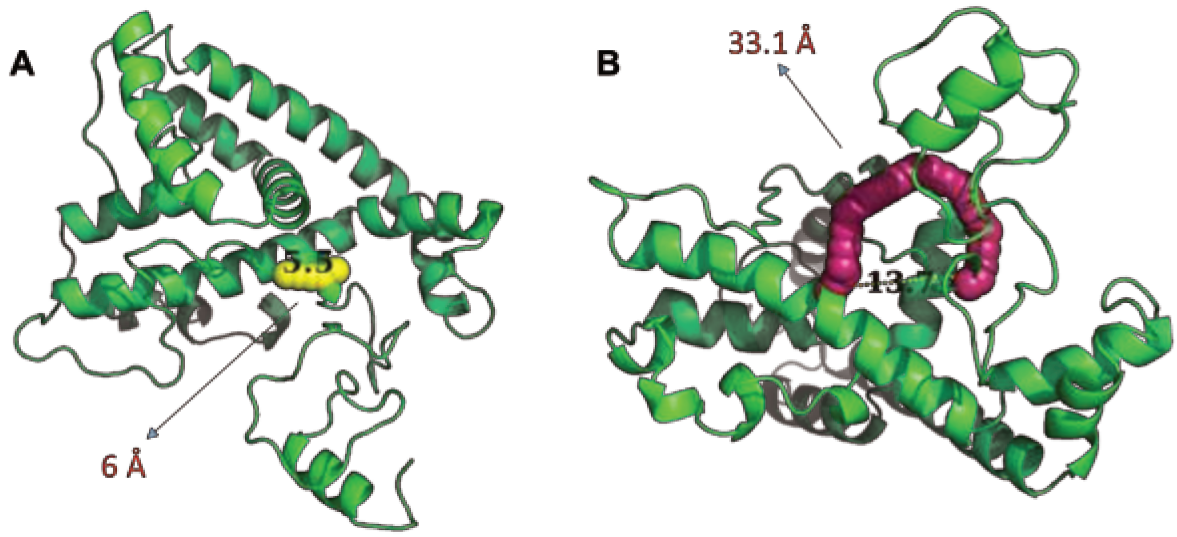


Figura 24: Representação das diferenças existentes entre a distância euclidiana (valores em preto) e a distância topológica (valores em vermelho). A distância entre dois pontos de uma proteína é melhor descrita pelo caminho acessível ao solvente. Esse caminho pode ser pouco discrepante como mostrado em A, em que a diferença existente é de 0,5 Å, ou com diferenças muito acentuadas, como em B, em que o valor da distância topológica representa mais que o dobro do valor da distância euclidiana.

Os modelos do conjunto **C** foram avaliados quanto a distância topológica entre os resíduos envolvendo as restrições experimentais obtidas, mantendo-se o valor

limite em 30 Å entre os C_α. A **Figura 25** indica a fração dos500 modelos de menor score dentre os modelos que respeitavam as distâncias euclidianas (veja seção 4.4) que respeitam cada distância topológica individualmente. A análise comparativa da **Figura 22** permite concluir que as restrições de curta distância na sequência tornam-se importantes na seleção quando a distância topológica é utilizada e, principalmente, as restrições de longa distância possuem um efeito discriminatório mais relevante na seleção de modelos, sendo que uma única restrição é capaz, em alguns casos, de reduzir para menos de 10% o número de candidatos.

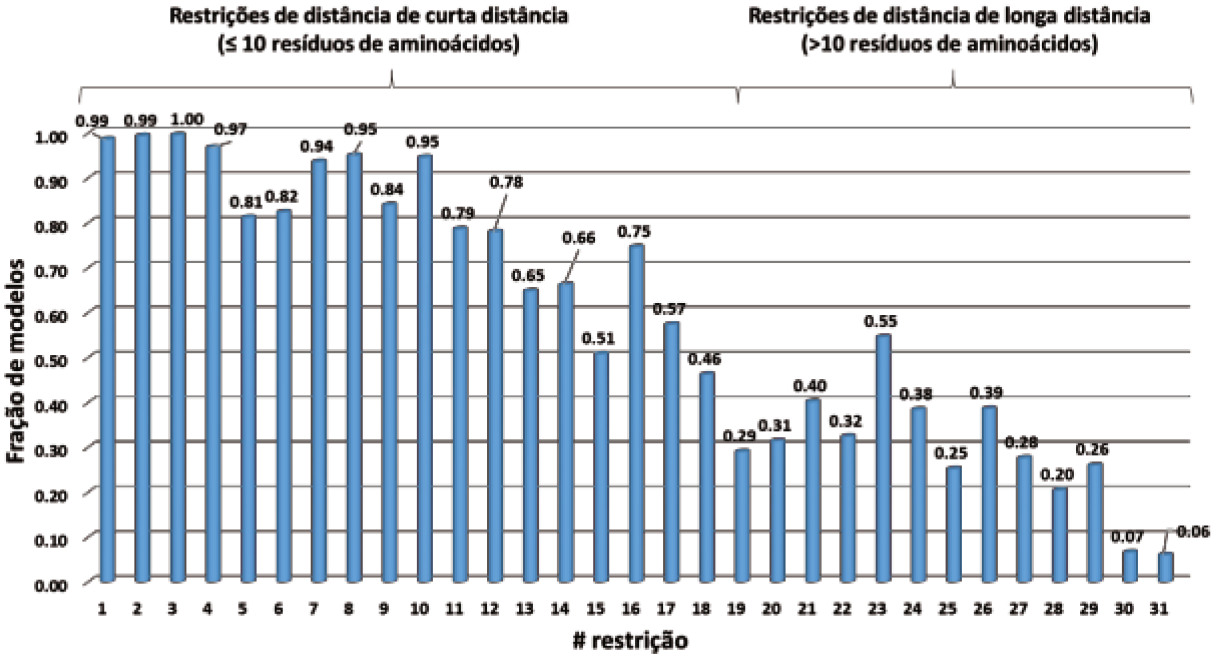


Figura 25: Relação entre o número de modelos filtrados por restrição utilizando-se a distância topológica. Todos os 500 modelos utilizados nessa análise possuíam as 31 restrições com distância euclidiana menores ou iguais a 30 Å. Restrições de curta e longa distância são importantes na discriminação entre modelos.

A influência do número de restrições na seleção de modelos utilizando como parâmetro de avaliação a distância topológica está apresentado na **Figura 26**. Assim como observado para as distâncias euclidianas, há uma tendência exponencial na discriminação dos modelos em relação ao número de restrições utilizadas, porém de forma mais acentuada, sendo que somente cinco modelos possuem simultaneamente 27 restrições abaixo de 30 Å.

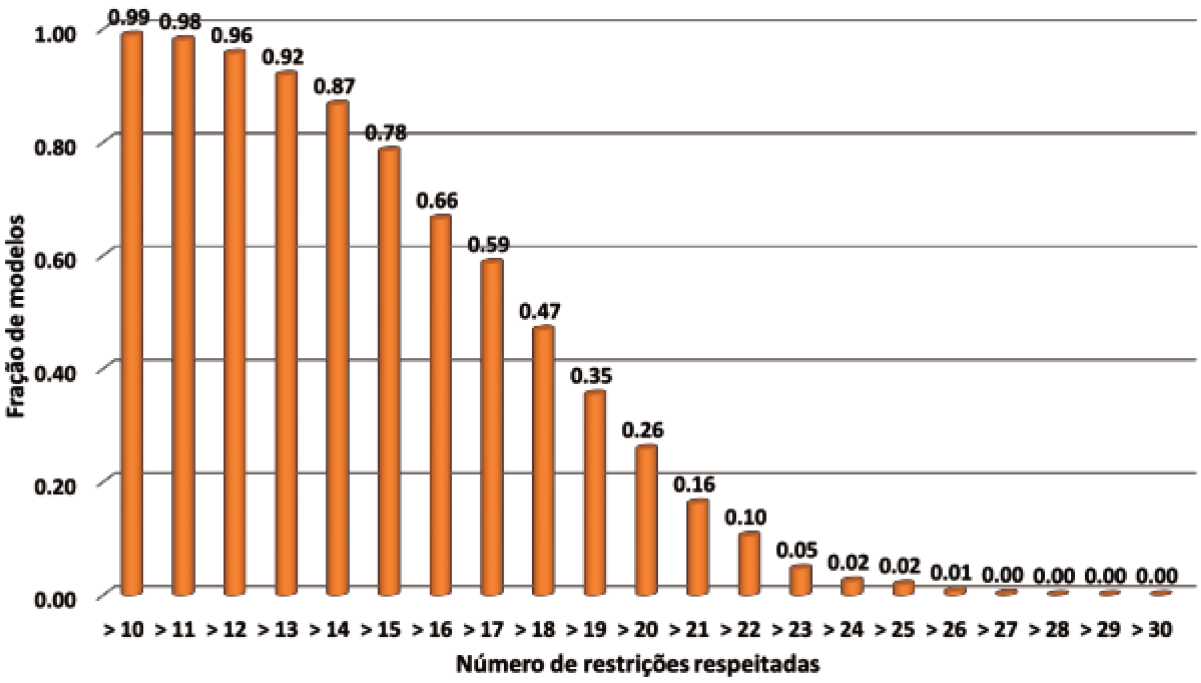


Figura 26: Influência do número de restrições experimentais na discriminação de modelos utilizando a distância topológica. Todos os 500 modelos utilizados nessa análise possuíam as 31 restrições com distância euclidiana menores ou iguais a 30 Å.

Em vista desses dados, seria necessário um estudo sistemático para se estabelecer quais os valores limites que devem ser adotados tendo em vista a dinâmica da proteína. Infelizmente, tal análise por dinâmica molecular é muito difícil, pois a demanda computacional para os cálculos da distância topológica é muito superior àquela necessária para o cálculo da distância euclidiana.

4.5 Agrupamentos de estruturas

A fim de verificar dentre os conjuntos de estruturas geradas quantos grupos distintos de estruturas realmente existiam, foram realizadas algumas análises de agrupamento por similaridade estrutural. Para isso, foram selecionadas as 500 estruturas que possuíam o menor *score* pelo Rosetta dentre aquelas que respeitaram todas as restrições de distância, dos conjuntos **B**, **C** e **D**. Este *score* possui uma correlação com a energia associada ao modelo e, ainda que seja uma medida aproximada, é razoável excluir modelos que possuem *scores* muito altos, como aqueles à direita da **Figura 27**.

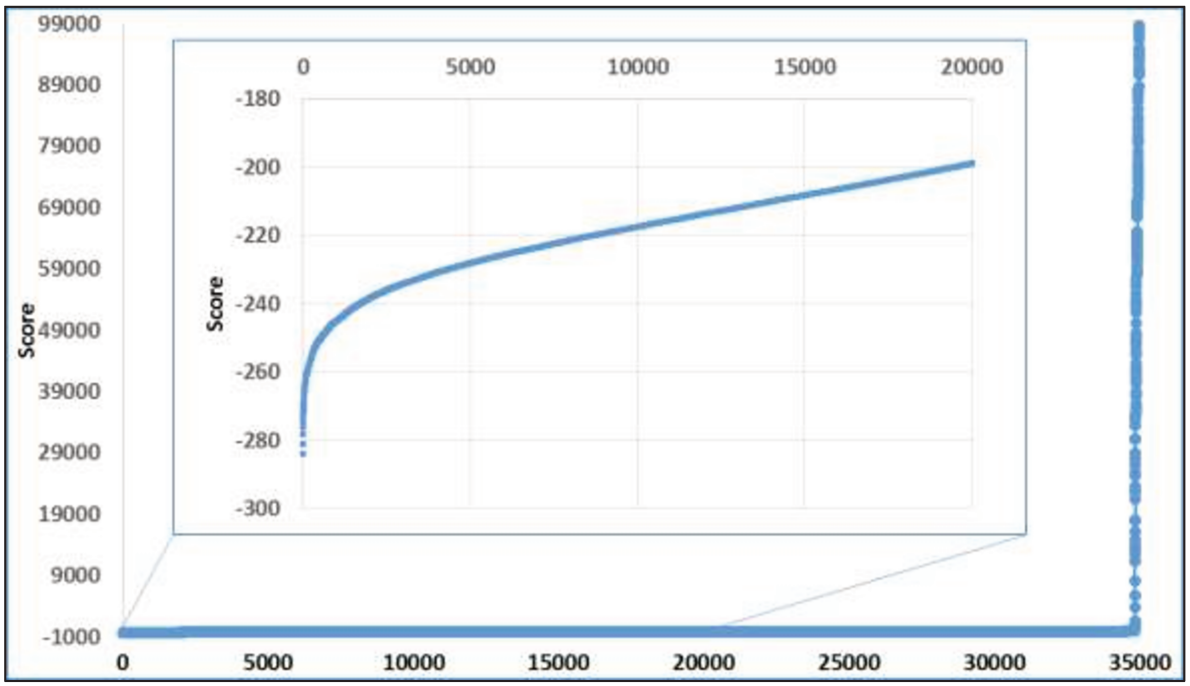


Figura 27: Distribuição de *scores* para modelos gerados pelo Rosetta. Estruturas que representam modelos mais razoáveis para a STC1 devem estar relacionadas a pontos à esquerda deste gráfico.

O programa MaxCluster é uma ferramenta que permite a comparação de estruturas de proteínas. A característica chave desse programa é sua habilidade de processar centenas de estruturas, seja com referência a uma única estrutura ou em uma comparação todos contra todos. Para realizar essas comparações inicialmente é realizado o alinhamento dessas estruturas, buscando-se o menor RMSD entre elas para, em seguida, se acessar através de um valor numérico a similaridade entre cada par de estruturas comparadas. Esse valor pode ser o já calculado RMSD entre as duas estruturas ou algum *score* que avalie o alinhamento estrutural. Dentre os scores disponíveis estão o GDT *score* e o MaxSub *score*.

O GDT (*Global Distance Test)*[88] é o método é utilizado no CASP a fim de avaliar a qualidade dos modelos gerados para os alvos propostos. O *score* é definido como:

$$GDT\ score = 100 * \frac{(C1 + C2 + C3 + C4)}{4N}$$

Onde:

C1 = Número de pares de resíduos superpostos abaixo de um quarto do limiar;

C2 = Número de pares de resíduos superpostos abaixo de metade do limiar;

C3 = Número de pares de resíduos superpostos abaixo do limiar;

C4 = Número de pares de resíduos superpostos abaixo de duas vezes o limiar.

O GDT score possui valores que variam entre 0 e 100, em que uma superposição aleatória entre estruturas não relacionadas terão um score entre 10 e 20. O limiar utilizado como padrão pelo programa é 4 Å. Dessa forma, um par de estruturas que possuam todos os pares de resíduos superpostos abaixo de 1 Å possuiria um GDT score de 100; um par de estruturas com todos os pares de resíduos superpostos entre 3-4 Å possuiria um GTD score de 25, e assim sucessivamente.

O MaxSub (*Maximal subset*)[89] é um algoritmo de busca que permite identificar o subconjunto máximo de resíduos pareados que podem ser sobrepostos dentro de uma dada distância limite, *d*. Esse subconjunto é avaliado segundo o MaxSub *score*. Após alinhamento das estruturas, as distâncias entre os resíduos pareados são calculadas e quaisquer resíduos que estejam dentro do limite da distância *d* definida é adicionado a um subconjunto. O MaxSub *score* é calculado para esse subconjunto como:

$$MaxSub = \frac{1}{N}\sum_{i}^{M}\frac{1}{(1+\frac{d_i^2}{d^2})}$$

Onde:

d_i = distância entre resíduos idênticos i

d = distância limite

M = O número de resíduos no subconjunto

N = O número de resíduos no modelo

O MaxSub *score* varia entre 0 e 1, onde 1 significa que as estruturas comparadas são idênticas. O MaxSub *score* original utiliza uma distância limite de 3,5 Å. Além disso, o score é calculado utilizando somente pares dentro do subconjunto M embora seja normalizado pelo tamanho N do modelo utilizado. Consequentemente

(a) o score é dependente do tamanho do modelo e (b) o score será o mesmo para dois modelos com estruturas idênticas dentro do subconjunto M mas que sejam diferentes fora desse subconjunto.

Após gerar a matriz de valores de comparação entre os modelos, as estruturas foram agrupadas utilizando o algoritmo do agrupamento do vizinho mais próximo restrito[90] (Restricted *Nearest Neighbour clustering*) presente no programa. Esse algoritmo constrói grupos utilizando a distância entre uma única estrutura, o centroide, e todas as demais. Dois modelos são considerados parte do mesmo grupo se eles estão próximos de um limite de corte, diz-se que eles são "vizinhos próximos". O modelo com maior número de vizinhos é escolhido como o centroide do grupo e, de acordo com um tamanho máximo especificado para cada grupo, os vizinhos mais próximos são retirados do universo de modelos juntamente com o centroide. Esse processo é repetido até que não haja mais grupos para serem formados dentro dos parâmetros estabelecidos – tamanho mínimo do grupo, ou limite de distância entre duas estruturas.

A fim de avaliar como a utilização de cada variável de comparação – RMSD, GDT *score* e MaxSub *score* – afetava o agrupamento e qual a qualidade da comparação entre os modelos ditos similares, foram realizados alguns testes com o conjunto de modelos ***C***.

Os parâmetros envolvendo cada variável de comparação foram ajustados para se agrupar pelo menos metade dos modelos do conjunto escolhido (pelo menos 250 estruturas) baseados em análises preliminares.

A **Figura 28** mostra a distribuição do *score* dos 500 modelos em relação ao RMSD para a estrutura de menor *score* desse conjunto. Em destaque aparecem pontos que mostram os centroides para cada variável de comparação. Ainda que eventualmente existam pontos de coincidência entre os centroides, fica claro que a agrupamento de modelos é criticamente dependente da variável que avalia o alinhamento entre as estruturas. Dentre as várias tentativas realizadas, percebeu-se que a variação dos parâmetros dentro de cada variável pouco influenciava na atribuição dos centroides dessa variável, pelo menos para os grupos de maior tamanho.

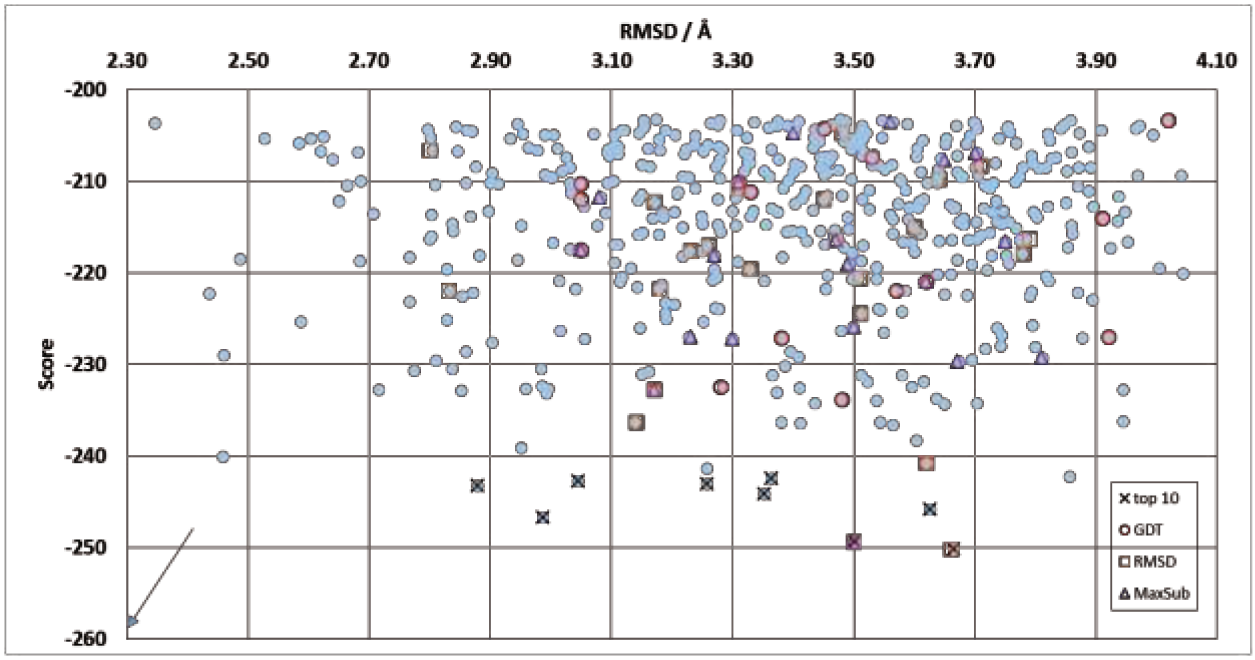


Figura 28: Distribuição dos *scores* dos modelos C em função do RMSD em relação ao modelo de menor *score* (ponto omitido; valor de *score* indicado pela seta). São destacadas as estruturas indicadas como centroides nos agrupamentos realizados e as estruturas com os dez menores *score* (legenda). Não existe na maior parte dos casos uma sobreposição de escolhas para os centroides dessas estruturas dentre as diferentes variáveis de agrupamento.

Os modelos do maior grupo de cada um dos três resultados foram analisados de forma mais detalhada. De posse das estruturas de um agrupamento é possível calcular como as outras variáveis classificam o alinhamento estrutural em relação ao centroide deste grupo.

O GDT *score* resultou no agrupamento de 255 modelos em 18 grupos, sendo o maior grupo contido por 75 estruturas. Os gráficos de correlação entre as três variáveis são mostrados na **Figura 29A-C**. Observa-se que a distribuição dos outros *scores* em relação ao RMSD é dispersa enquanto que existe uma correlação razoável entre o GDT *score* e o MaxSub *score*. A análise manual comparando-se as estruturas de maior e menor GDT *score* dentro do grupo em relação ao centroide – **Figuras 29D-E**, respectivamente – mostrou que esta variável é eficiente no agrupamento de estruturas similares.

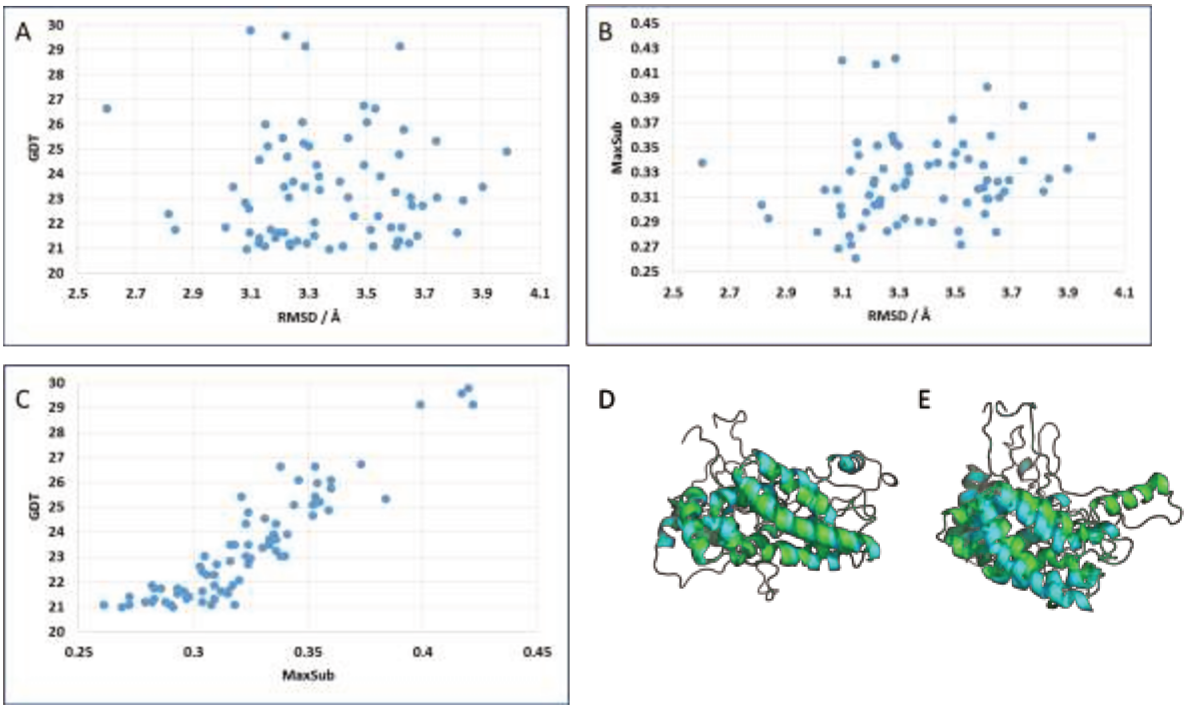


Figura 29: Avaliação do GDT *score* no agrupamento das estruturas. A relação entre as variáveis considerando o maior grupo é mostrada nos gráficos (A), (B) e (C). O *GDT score* e o MaxSub *score* possui uma correlação evidente, enquanto que o RMSD possui uma distribuição dispersa frente as outras variáveis. As Figuras (D) e (E) indicam que as estruturas possuem correlação estrutural dentro de um mesmo grupo.

A utilização do MaxSub score resultou no agrupamento de 371 estruturas em 19 grupos, sendo que o maior grupo apresentava 60 estruturas. As tendências encontradas neste agrupamento são mostradas nas **Figuras 30A-C**. Observa-se os mesmos perfis de correlação que foram encontradas no agrupamento utilizando o GDT *score,* sendo que a correlação existente entre este *score* e o MaxSub *score* é ainda mais acentuada. As **Figuras 30D-E** mostram o alinhamento estrutural entre o centroide e os modelos com maior e menor *score* dentro do maior grupo, respectivamente. O alinhamento relacionado com a estrutura de menor *score* possui um GDT *score* de 18, indicando, segundo esse parâmetro, que se trataria de um alinhamento aleatório e que esses modelos não possuem uma boa correlação estrutural. Apesar de a análise em outros grupos indicar que as estruturas de menor *score* possuem visualmente um alinhamento pobre em relação ao centroide, a utilização desse parâmetro resultou em 40% mais estruturas agrupadas quando

comparada ao GDT *score*, sendo assim a sua utilização pode incorrer em pequenas falhas formando mais grupos ou ao atribuir mais estruturas por grupo.

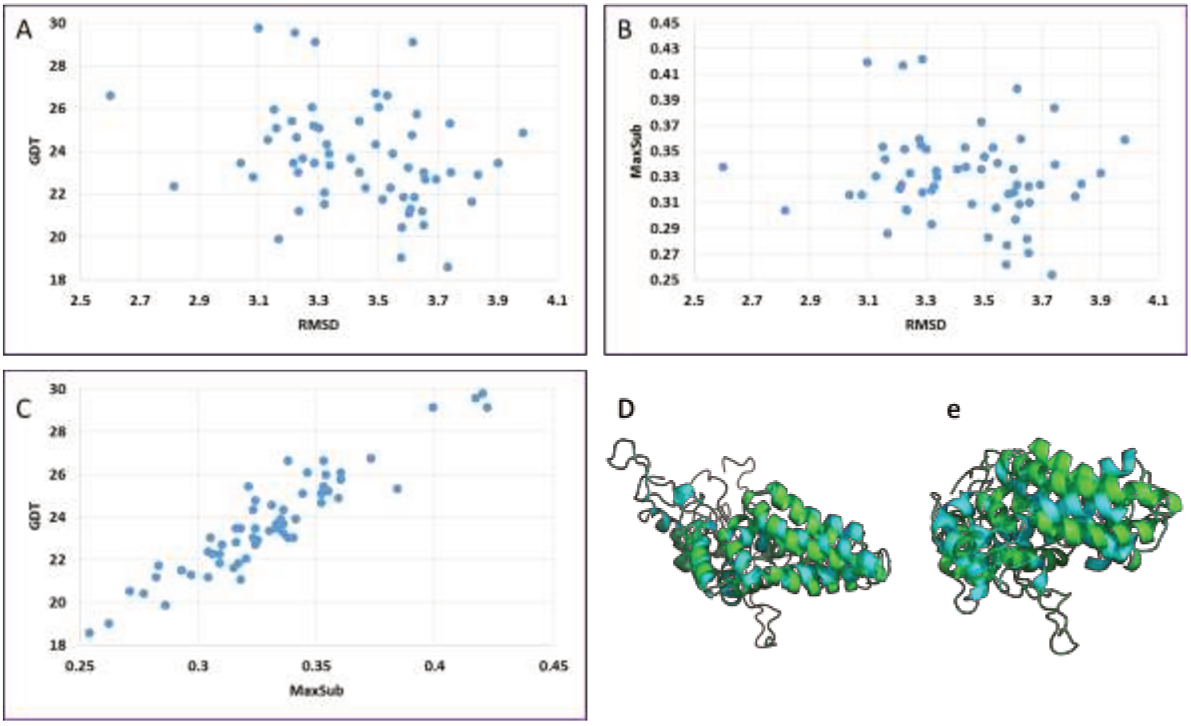


Figura 30: Avaliação do MaxSub *score* no agrupamento de estruturas. A relação entre as variáveis considerando o maior grupo é mostrada nos gráficos (A), A(B) e (C). O *GDT score* e o MaxSub *score* possuem uma correlação evidente, enquanto que o RMSD possui uma distribuição dispersa frente as outras variáveis. As Figuras (D) e (E) indicam que as estruturas possuem correlação estrutural dentro de um mesmo grupo porém estruturas com menor *score* possuem qualidade inferior de similaridade quando comparada ao GDT *score*.

O RMSD resultou em um agrupamento de 254 estruturas em 23 grupos, sendo o maior grupo constituído de 44 estruturas. A **Figura 31A-C** mostra que o padrão de correlação é o mesmo indicado anteriormente, destacando-se que boa parte das estruturas possuem um GDT *score* abaixo de 20 e um MaxSub *score* inferior a 0,25. A análise manual das estruturas com menor e maior RMSD dentro do grupo em relação ao centroide - **Figura 31D-E**, respectivamente - permite confirmar que essa variável não é efetiva para realizar o agrupamento, sendo que a atribuição de similaridade pode ser considerada aleatória entre estruturas para um mesmo grupo.

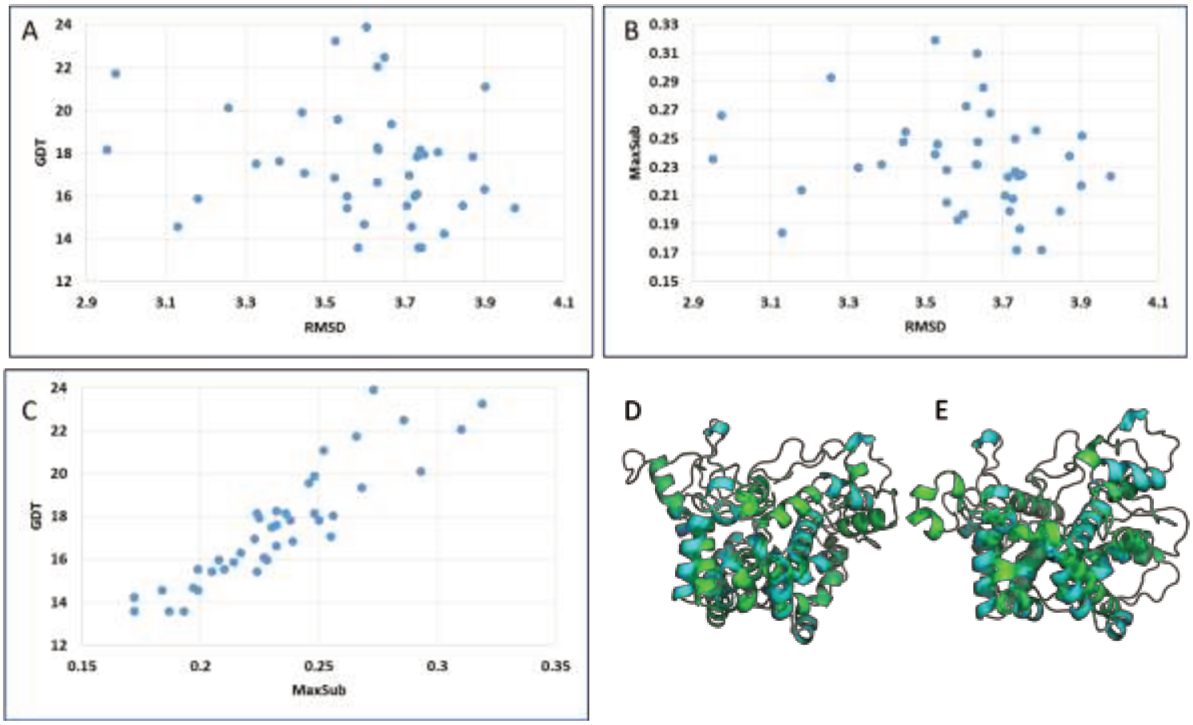


Figura 31: Avaliação do RMSD no agrupamento de estruturas. A relação entre as variáveis considerando o maior grupo é mostrada nos gráficos (A), (B) e (C). O *GDT score* e o MaxSub *score* possuem uma correlação evidente, enquanto que o RMSD possui uma distribuição dispersa frente as outras variáveis. As Figuras (D) e (E) indicam que as estruturas possuem uma baixa similaridade, podendo ser consideradas essencialmente diferentes.

Como o RMSD é um pré-requisito de qualquer análise para o alinhamento, a sua utilização como valor de comparação requer um tempo computacional muito inferior quando comparado com a utilização do GDT *score* ou o MaxSub *score*, o que pode ser útil quando se quer obter informações mais grosseiras de comparação global com um conjunto muito grande de estruturas.

Optou-se pelos modelos agrupados pelo GDT *score*, de forma a obter grupos possuíssem estruturas com alto grau de semelhança. Em resumo no fim das análises haviam nove, 18 e 13 grupos dos conjuntos **B**, **C** e **D**, respectivamente. Os centroides de cada grupo (40 estruturas) foram analisados quanto a sua similaridade utilizando-se os mesmos parâmetros com o GDT *score*. As estruturas foram agrupadas em 20 estruturas singulares e mais dois conjuntos com 15 e cinco estruturas. Dessa forma, obteve-se ao final 22 candidatos estruturalmente distintos para a STC1. A **Figura 32** mostra um desses candidatos com destaque para alguns resíduos envolvidos nas

restrições de distância, mostrando que essas estruturas são compatíveis com as restrições de distâncias e potenciais modelos para a estrutura da STC1.

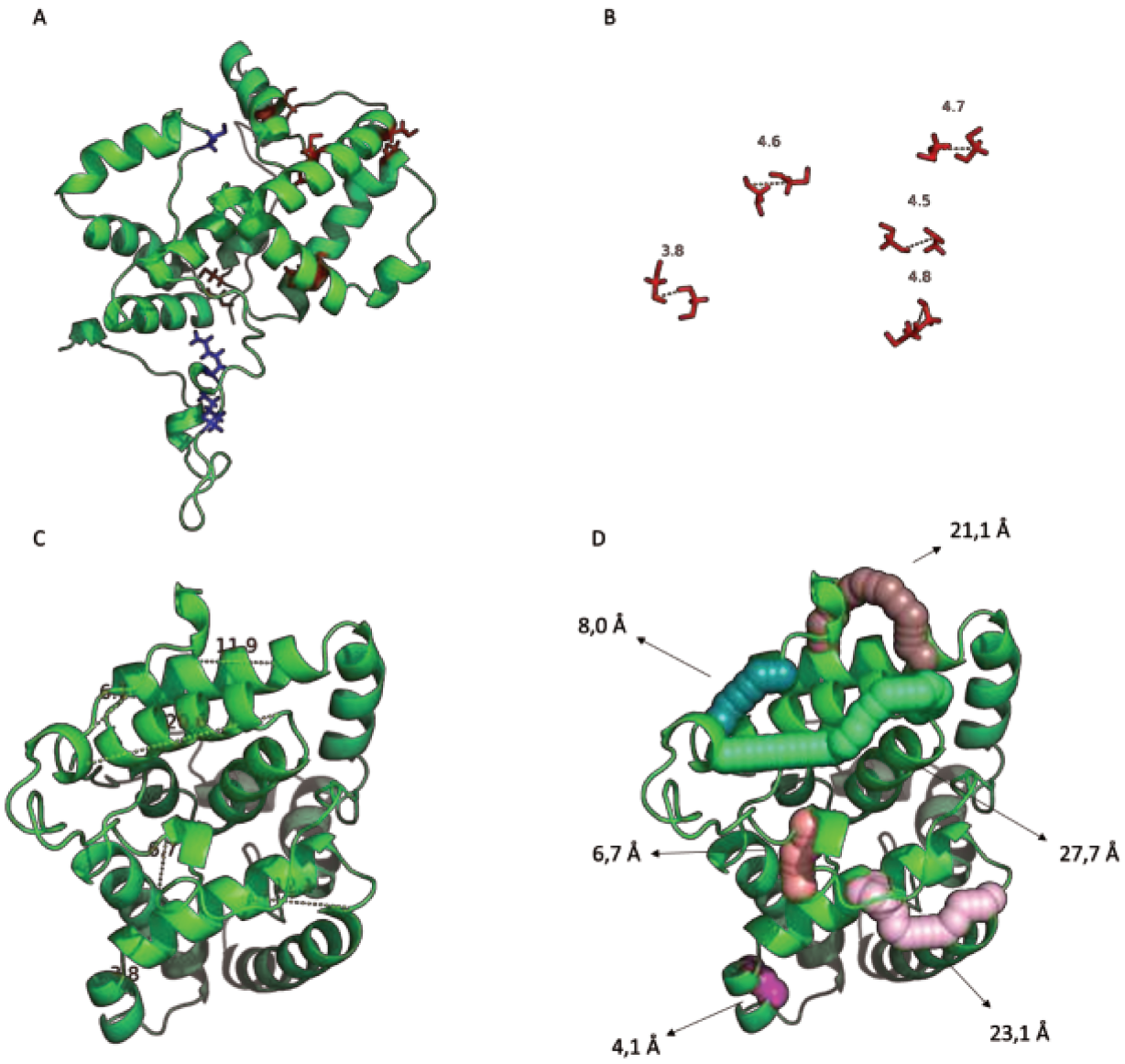


Figura 31: Representação de uma estrutura candidata para a STC1-HT. (A) Estrutura com destaque em vermelho para as cisteínas envolvidas nas ligações dissulfeto intra-cadeia; em azul, resíduos encontrados que estão envolvidas na dimerização. (B) Cisteínas envolvidas nas ligações de dissulfeto com os valores das distâncias euclidianas Cα-Cα destacadas; (B) Destaque para cinco ligações cruzadas encontradas com as distâncias euclidianas Cα-Cα destacadas; (D) As mesmas ligações cruzadas são mostradas de acordo com o caminho acessível ao solvente; cada seta indica o valor da distância topológica associada àquele caminho.

5. Conclusões

A técnica de ligação cruzada acoplada a espectrometria de massas permitiu identificar 42 restrições de distância para a STC1-HT, sendo 11 destas espécies que envolviam o N-terminal.

Foi possível identificar uma espécie de ligação cruzada simétrica na porção C-terminal, comprovando a região de interação do dímero.

Dentre os programas de modelagem avaliados, o programa Rosetta foi o programa que gerou estruturas mais condizentes com as restrições de distância, sendo a ausência da predição de estrutura secundária foi fundamental para isso.

O número de restrições foi uma ferramenta importante para a geração e a seleção de modelos, sendo que restrições de longo alcance são mais eficientes na discriminação de modelos. Além disso, a distância topológica possui maior eficiência de discriminação em relação a distância euclidiana, sendo ainda necessário um estudo sistemático para a sua implementação quanto aos valores para o DSS.

As simulações de dinâmica molecular para algumas estruturas indicam que a porção C-terminal no monômero possui maior flexibilidade, o que corrobora alguns dos dados de ligação cruzada envolvendo essa região. Entretanto, não indicam uma solução para os dados de restrição de distância envolvendo a porção N-terminal da STC1-HT.

Dentre as variáveis disponíveis no programa MaxCluster, o GDT *score* mostrou maior eficiência em agrupar estruturas similares, possuindo uma correlação grande com o MaxSub *score*. Já RMSD é uma variável que proporciona muitos encaixes aleatórios e agrupa modelos estruturalmente não correlacionados.

Foram selecionados 22 candidatos estruturalmente distintos para a STC1 dentre 1.500 modelos de menor *score* gerados pelo Rosetta.

De forma geral, este trabalho é o primeiro estudo a investigar a viabilidade do uso de restrições de distância obtidas por experimentos de ligação cruzada no estudo da estrutura terciária de proteínas. Os métodos aqui desenvolvidos mostraram quais variáveis são importantes na determinação de modelos e determinaram a forma mais

eficiente até o momento de se utilizar esses dados experimentais de restrições de distância na modelagem da estrutura terciária da STC1. Uma avaliação desses modelos finais frente aos dados experimentais de cristalografia de proteínas constitui uma perspectiva deste estudo.

6. Referências Bibliográficas

1. Nelson, D. L. & Cox, M. M. *Lehninger Principles of Biochemistry*. (W. H. Freeman, 2004).
2. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6,** 197–208 (2005).
3. Cheng, J. & Baldi, P. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics* **22,** 1456–1463 (2006).
4. Wu, S. & Zhang, Y. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins* **72,** 547–556 (2008).
5. Wu, S. & Zhang, Y. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res.* **35,** 3375–3382 (2007).
6. Ołdziej, S. *et al.* Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 7547–7552 (2005).
7. Hagler, A. T., Huler, E. & Lifson, S. Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *J. Am. Chem. Soc.* **96,** 5319–5327 (1974).
8. Weiner, S. J. *et al.* A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106,** 765–784 (1984).
9. MacKerell, A. D. *et al.* All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **102,** 3586–3616 (1998).
10. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **118,** 11225–11236 (1996).
11. Skolnick, J. In quest of an empirical potential for protein structure prediction. *Curr. Opin. Struct. Biol.* **16,** 166–171 (2006).
12. Zhang, Y., Kolinski, A. & Skolnick, J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys. J.* **85,** 1145–1164 (2003).
13. Zhang, Y., Hubner, I. A., Arakaki, A. K., Shakhnovich, E. & Skolnick, J. On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 2605–2610 (2006).
14. Kim, D. E., Chivian, D. & Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **32,** W526–531 (2004).
15. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5,** 725–738 (2010).
16. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12,** 7–8 (2015).
17. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9,** 40 (2008).
18. Xu, D. & Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80,** 1715–1735 (2012).
19. Das, R. & Baker, D. in *Annual Review of Biochemistry* **77,** 363–382 (Annual Reviews, 2008).
20. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using rosetta. *Numer. Comput. Methods Pt D* **383,** 66–+ (2004).
21. Baker, D. Protein folding, structure prediction and design. *Biochem. Soc. Trans.* **42,** 225–229 (2014).
22. Dos Reis, M. A., Aparicio, R. & Zhang, Y. Improving protein template recognition by using small-angle x-ray scattering profiles. *Biophys. J.* **101,** 2770–2781 (2011).
23. Russel, D. *et al.* Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* **10,** e1001244 (2012).
24. Schneidman-Duhovny, D. *et al.* A method for integrative structure determination of protein-protein complexes. *Bioinforma. Oxf. Engl.* **28,** 3282–3289 (2012).

25. Baker, J. & Ando, D. J. *Mass Spectrometry: Analytical Chemistry by Open Learning*. (Wiley, 1999).
26. Munson, M. S. B. & Field, F. H. Chemical Ionization Mass Spectrometry. I. General Introduction. *J. Am. Chem. Soc.* **88,** 2621–2630 (1966).
27. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71 (1989).
28. Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.* **60,** 2299–2301 (1988).
29. Kinter, M. & Sherman, N. E. *Protein Sequencing and Identification Using Tandem Mass Spectrometry*. (John Wiley & Sons, 2000).
30. Hu, Q. *et al.* The Orbitrap: a new mass spectrometer. *J. Mass Spectrom. JMS* **40,** 430–443 (2005).
31. Aebersold, R. & Goodlett, D. R. Mass spectrometry in proteomics. *Chem. Rev.* **101,** 269–295 (2001).
32. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).
33. Patterson, S. D. & Aebersold, R. H. Proteomics: the first decade and beyond. *Nat. Genet.* **33 Suppl,** 311–323 (2003).
34. Steen, H. & Mann, M. The abc's (and xyz's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5,** 699–711 (2004).
35. Yates, J. R. Mass spectral analysis in proteomics. *Annu. Rev. Biophys. Biomol. Struct.* **33,** 297–316 (2004).
36. Domon, B. & Aebersold, R. Mass Spectrometry and Protein Analysis. *Science* **312,** 212–217 (2006).
37. Cravatt, B. F., Simon, G. M. & Yates Iii, J. R. The biological impact of mass-spectrometry-based proteomics. *Nature* **450,** 991–1000 (2007).
38. Wong, S. S. *Chemistry of Protein Conjugation and Cross-Linking*. (CRC Press, 1991).
39. Sinz, A. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. *Mass Spectrom. Rev.* **25,** 663–682 (2006).
40. Sinz, A. Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. *J. Mass Spectrom. JMS* **38,** 1225–1237 (2003).
41. Young, M. M. *et al.* High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 5802–5806 (2000).
42. Singh, P., Panchaud, A. & Goodlett, D. R. Chemical cross-linking and mass spectrometry as a low-resolution protein structure determination technique. *Anal. Chem.* **82,** 2636–2642 (2010).
43. Sinz, A. Isotope-labeled photoaffinity reagents and mass spectrometry to identify protein-ligand interactions. *Angew. Chem. Int. Ed Engl.* **46,** 660–662 (2007).
44. Schmidt, A., Kalkhof, S., Ihling, C., Cooper, D. M. F. & Sinz, A. Mapping protein interfaces by chemical cross-linking and Fourier transform ion cyclotron resonance mass spectrometry: application to a calmodulin / adenylyl cyclase 8 peptide complex. *Eur. J. Mass Spectrom. Chichester Engl.* **11,** 525–534 (2005).
45. Alley, S. C., Ishmael, F. T., Jones, A. D. & Benkovic, S. J. Mapping Protein−Protein Interactions in the Bacteriophage T4 DNA Polymerase Holoenzyme Using a Novel Trifunctional Photo-cross-linking and Affinity Reagent. *J. Am. Chem. Soc.* **122,** 6126–6127 (2000).
46. Trester-Zedlitz, M. *et al.* A Modular Cross-Linking Approach for Exploring Protein Interactions. *J. Am. Chem. Soc.* **125,** 2416–2425 (2003).
47. Tang, X., Munske, G. R., Siems, W. F. & Bruce, J. E. Mass Spectrometry Identifiable Cross-Linking Strategy for Studying Protein−Protein Interactions. *Anal. Chem.* **77,** 311–318 (2004).
48. Fioramonte, M. *et al.* Analysis of secondary structure in proteins by chemical cross-linking coupled to MS. *Proteomics* **12,** 2746–2752 (2012).

49. Pereira, M. B. M. *et al.* αB-crystallin interacts with and prevents stress-activated proteolysis of focal adhesion kinase by calpain in cardiomyocytes. *Nat. Commun.* **5,** 5159 (2014).
50. Santos, A. M. *et al.* FERM domain interaction with myosin negatively regulates FAK in cardiomyocyte hypertrophy. *Nat. Chem. Biol.* **8,** 102–110 (2012).
51. Tiroli-Cepeda, A. O., Lima, T. B., Balbuena, T. S., Gozzo, F. C. & Ramos, C. H. I. Structural and functional characterization of the chaperone Hsp70 from sugarcane. Insights into conformational changes during cycling from cross-linking/mass spectrometry assays. *J. Proteomics* **104,** 48–56 (2014).
52. Iglesias, A. H., Santos, L. F. A. & Gozzo, F. C. Identification of Cross-Linked Peptides by High-Resolution Precursor Ion Scan. *Anal. Chem.* **82,** 909–916 (2010).
53. Lima, D. B. *et al.* SIM-XL: A powerful and user-friendly tool for peptide cross-linking analysis. *J. Proteomics* doi:10.1016/j.jprot.2015.01.013
54. Wagner, G. F. & Dimattia, G. E. The stanniocalcin family of proteins. *J. Exp. Zoolog. A Comp. Exp. Biol.* **305,** 769–780 (2006).
55. Tanega, C., Radman, D. P., Flowers, B., Sterba, T. & Wagner, G. F. Evidence for stanniocalcin and a related receptor in annelids. *Peptides* **25,** 1671–1679 (2004).
56. Ishibashi, K. & Imai, M. Prospect of a stanniocalcin endocrine/paracrine system in mammals. *Am. J. Physiol. Renal Physiol.* **282,** F367–375 (2002).
57. Yoshiko, Y., Maeda, N. & Aubin, J. E. Stanniocalcin 1 stimulates osteoblast differentiation in rat calvaria cell cultures. *Endocrinology* **144,** 4134–4143 (2003).
58. Worthington, R. A. *et al.* Expression and localisation of stanniocalcin 1 in rat bladder, kidney and ovary. *Electrophoresis* **20,** 2071–2076 (1999).
59. Chang, A. C.-M., Jellinek, D. A. & Reddel, R. R. Mammalian stanniocalcins and cancer. *Endocr. Relat. Cancer* **10,** 359–373 (2003).
60. Ismail, R. S. *et al.* Differential gene expression between normal and tumor-derived ovarian epithelial cells. *Cancer Res.* **60,** 6744–6749 (2000).
61. Yeung, H. Y. *et al.* Hypoxia-inducible factor-1-mediated activation of stanniocalcin-1 in human cancer cells. *Endocrinology* **146,** 4951–4960 (2005).
62. Wascher, R. A. *et al.* Stanniocalcin-1: a novel molecular blood and bone marrow marker for human breast cancer. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **9,** 1427–1435 (2003).
63. Dos Santos, M. T. *et al.* Human stanniocalcin-1 interacts with nuclear and cytoplasmic proteins and acts as a SUMO E3 ligase. *Mol. Biosyst.* **7,** 180–193 (2011).
64. Geiss-Friedlander, R. & Melchior, F. Concepts in sumoylation: a decade on. *Nat. Rev. Mol. Cell Biol.* **8,** 947–956 (2007).
65. Hickey, C. M., Wilson, N. R. & Hochstrasser, M. Function and regulation of SUMO proteases. *Nat. Rev. Mol. Cell Biol.* **13,** 755–766 (2012).
66. Chang, A. C.-M., Dunham, M. A., Jeffrey, K. J. & Reddel, R. R. Molecular cloning and characterization of mouse stanniocalcin cDNA. *Mol. Cell. Endocrinol.* **124,** 185–187 (1996).
67. Gerritsen, M. E. & Wagner, G. F. Stanniocalcin: no longer just a fish tale. *Vitam. Horm.* **70,** 105–135 (2005).
68. Trindade, D. M., Silva, J. C., Navarro, M. S., Torriani, I. C. & Kobarg, J. Low-resolution structural studies of human Stanniocalcin-1. *BMC Struct. Biol.* **9,** 57 (2009).
69. Park, C. Y., Käll, L., Klammer, A. A., MacCoss, M. J. & Noble, W. S. Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.* **7,** 3022–3027 (2008).
70. Yang, B. *et al.* Identification of cross-linked peptides from complex samples. *Nat. Methods* **9,** 904–906 (2012).
71. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9,** 40 (2008).
72. Xu, D. & Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80,** 1715–1735 (2012).
73. Das, R. & Baker, D. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **77,** 363–382 (2008).

74. Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H. & Meiler, J. Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You. *Biochemistry (Mosc.)* **49,** 2987–2998 (2010).
75. Raman, S. *et al.* Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins Struct. Funct. Bioinforma.* **77,** 89–99 (2009).
76. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292,** 195–202 (1999).
77. Kim, D. E., Chivian, D. & Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **32,** W526–531 (2004).
78. Kahraman, A., Malmstroem, L. & Aebersold, R. Xwalk: computing and visualizing distances in cross-linking experiments. *Bioinformatics* **27,** 2163–2164 (2011).
79. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26,** 1781–1802 (2005).
80. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph. Model.* **14,** 33–38 (1996).
81. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79,** 926–935 (1983).
82. Martínez, J. M. & Martínez, L. Packing optimization for automated generation of complex system's initial configurations for molecular dynamics and docking. *J. Comput. Chem.* **24,** 819–825 (2003).
83. Martinez, L., Andrade, R., Birgin, E. G. & Martinez, J. M. PACKMOL: A Package for Building Initial Configurations for Molecular Dynamics Simulations. *J. Comput. Chem.* **30,** 2157–2164 (2009).
84. MATLAB and Statistics Toolbox Release 2014b, The MathWorks, Inc., Natick, Massachusetts, United States.
85. Fass, D. Disulfide Bonding in Protein Biophysics. *Annu. Rev. Biophys.* **41,** 63–79 (2012).
86. Merkley, E. D. *et al.* Distance restraints from crosslinking mass spectrometry: mining a molecular dynamics simulation database to evaluate lysine-lysine distances. *Protein Sci. Publ. Protein Soc.* **23,** 747–759 (2014).
87. Kahraman, A. *et al.* Cross-Link Guided Molecular Modeling with ROSETTA. *PLoS ONE* **8,** e73411 (2013).
88. Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* **31,** 3370–3374 (2003).
89. Siew, N., Elofsson, A., Rychlewski, L. & Fischer, D. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinforma. Oxf. Engl.* **16,** 776–785 (2000).
90. Shortle, D., Simons, K. T. & Baker, D. Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 11158–11162 (1998).