Universidade de Brasília – UnB
Instituto de Ciências Biológicas – IB
Departamento de Biologia Celular – CEL
Laboratório de Biofísica Teórica e Computacional – LBTC

Coevolução Molecular em Canais Iônicos e Neurotoxinas

Aluna: Camila Ferreira Thé Pontes
Orientador: PhD Werner Treptow

Dissertação apresentada ao Departamento de Biologia Celular
do Instituto de Ciências Biológicas da Universidade de Brasília
para obtenção do grau mestre em Biologia Molecular.

Brasília – Julho de 2016

Agradecimentos

Primeiramente, gostaria de agradecer ao Caio por ter me auxiliado com a parte computacional do projeto. Grande parte do que foi feito não teria sido possível sem a sua ajuda. Agradeço também ao Matheus (IC) por ter me ajudado com a montagem do banco de toxinas. Obrigada pela dedicação, pelo interesse e pelo compromisso. Espero que o aprendizado tenha sido proveitoso. Ao Miguel por ter me ajudado com as figuras. Ao meu orientador, prof. Werner Treptow, pela paciência e pelo estímulo. A todos os meus demais colegas de laboratório por terem me acolhido no grupo e me ajudado nos momentos de dificuldade. Enfim, agradeço a todos que contribuíram de alguma forma para a realização deste trabalho. Também à CAPES e ao CNPq pelo financiamento e à Universidade de Brasília pelo apoio.

Gostaria de agradecer ainda aos meus pais, que desde sempre me estimularam a estudar e a aprender. Se não fossem vocês, eu com certeza não estaria hoje realizando este trabalho. E, por fim, gostaria de agradecer ao meu namorado, Arthur, por ter me suportado em todos os momentos de estresse e correria. Muito, muito, muito obrigada pela paciência, pelo carinho e pela força que você me deu.

Gostaria ainda de ressaltar que tudo que fiz foi feito com muito interesse e dedicação. Fico muito feliz por ter tido a oportunidade de realizar este trabalho.

Comece fazendo o que é necessário, depois o que é possível e, de repente, você estará fazendo o impossível.

S. Francisco de Assis

Resumo

A peçonha de escorpiões contém diversos tipos de neurotoxinas que podem interagir entre si para modular a função de canais iônicos (Quintero-Hernández et al., 2013). A ação desses polipeptídeos leva à ativação de canais de sódio e inibição de canais de potássio, causando um elevado influxo de sódio e a liberação de neurotransmissores, seguida por um bloqueio da excitabilidade celular (Gurevitz et al., 2015). Apesar de possuírem estrutura 3D similar, as chamadas α- e β-toxinas de escorpião afetam canais iônicos de sódio dependentes de voltagem (Na_V) por meio de mecanismos diferentes: as α-toxinas interagem com o sítio 3 no domínio sensor de voltagem IV (VSD-IV) e inibem o processo de inativação rápida do canal (Quintero-Hernández et al., 2013), enquanto as β-toxinas interagem com o sítio 4 no VSD-II e causam a hiperativação do canal por meio de um mecanismo de aprisionamento do sensor de voltagem (Pedraza Escalona and Possani, 2013). Em um contexto evolutivo, espera-se que o sistema composto por esses dois tipos de toxinas e os seus alvos moleculares, os VSD de Na_V, tenham sofrido um processo de coevolução molecular. Partindo do princípio de que seja possível detectar, através da análise de sequências primárias, sinais de coevolução molecular que determinem a seletividade e afinidade entre os pares toxina-VSD, foi possível propor um um modelo evolutivo de interação e seletividade entre α- e β-toxinas de escorpião e VSD-II e -IV de Na_V, o qual representa o melhor conjunto possível de interações toxina-VSD. Para tanto, foi desenvolvido um algoritmo genético capaz de otimizar, baseado em um critério de energia e acoplamento, um dado sistema composto por dois conjuntos de posições de aminoácidos, obtidos de dois alinhamentos múltiplos de sequências (MSA) de proteínas. O algoritmo genético foi desenvolvido para encontrar a melhor forma de parear as sequências do MSA1 com as sequências do MSA2 de forma a minimizar a energia de interação total dos pares. O modelo otimizado de coevolução (MOC) apresentou dois grupos bem definidos, um formado por interações entre α-toxinas e VSD-IV e o outro composto por interações entre β-toxinas e VSD-II. Esse resultado indica que o algoritmo foi capaz de encontrar uma solução realista para o problema. O modelo obtido fornece informações importantes sobre quais interações entre resíduos definem as regras para as afinidade diferenciais entre β-toxinas e VSD-II, e α-toxinas e VSD-IV. Com isso, foi possível inferir um conjunto de resíduos que caracteriza a superfície funcional de cada grupo de toxinas. Os resultados obtidos são corroborados por resultados experimentais da literatura.

Palavras-chave: coevolução molecular, neurotoxinas, canais iônicos

Abstract

Scorpion venoms contain several types of neurotoxins that might interact with each other, modulating the function of ion channels (Quintero-Hernández et al., 2013). The action of these polipeptides leads to the activation of sodium channels and inhibition of potassium channels, causing a high sodium influx and the liberation of neurotransmitters, followed by a blockage of cellular excitability (Gurevitz et al., 2015). Regardless of their similar 3D structures, the so-called α- and β-scorpion toxins affect voltage-gated sodium channels (Na_V) through very different mechanisms: α-toxins interact with the extracellular site 3 in the voltage sensor domain IV (VSD-IV) and inhibit the rapid channel inactivation process (Quintero-Hernández et al., 2013), while β-toxins interact with site 4 in VSD-II and cause channel hyperactivation through a voltage sensor trapping mechanism (Pedraza Escalona and Possani, 2013). In an evolutionary context, it is expected that the system composed of this two types of gating modifier toxins plus the targeted Na_V VSD will present some coevolution traces. Starting from the hypotheses that it is possible to detect, through sole primary sequence analysis, signals of molecular coevolution determining selectivity and specificity between pairs of interacting proteins, it was possible to propose an evolutionary model of interaction and selectivity between scorpion α- and β-toxins and Na_V VSD-II and -IV, which represents the best possible arrangement of interacting pairs. To achieve that, a self-developed and implemented genetic algorithm that was able to optimize, based on an energy-coupling criterion, a given system composed of two sets of information channels coming from two different protein multiple sequence alignments (MSA) was used. Basically, the genetic algorithm was designed to find the best way of pairing the sequences coming from MSA1 with the sequences coming from MSA2 in order to minimize the overall interaction energy of the pairs. The optimized model presented two well-defined groups, one composed of α-toxins interacting with VSD-IV and the other composed of β-toxins interacting with VSD-II. This result indicates that the model is probably accurate. Going one step further, we applied PCA to extract important information from the optimized model about the interacting residues in the two groups (β-toxins, VSD-II and α-toxins, VSD-IV). It was then possible to infer the set of residues responsible for the unique features observed in the two groups of toxins. The results obtained in this last step are in conformation with data coming from experimental assays.

Keywords: molecular coevolution, neurotoxins, ion channels

Sumário

Lista de Figuras

Lista de Tabelas

Siglas e Abreviações

VGIC: *Voltage Gated Ion Channel*

VSD: *Voltage Sensor Domain*

Na^+: íon de sódio

K^+: íon de potássio

Na_V: canais iônicos dependentes de voltagem seletivos a sódio

hNa_V: canais Na_V de humanos

K_V: canais iônicos dependentes de voltagem seletivos a potássio

NaScTxs: toxinas de escorpião que agem em canais de sódio

MSA: *Multiple Sequence Alignment*

HMM: *Hidden Markov Model*

PDB: *Protein Data Bank*

UniProtKB: *Universal Protein Resource Knowledgebase*

TI: Teoria da Informação

AG: Algoritmo genético

CI: Canal de Informação

MOC: Modelo Otimizado de Coevolução

PCA: *Principal Component Analysis*

RMSD: *Root-Mean-Square Deviation*

SASA: *Solvent Accessible Surface Area*

Introdução

A Hipótese da Rainha Vermelha, proposta por van Valen em 1973 (Valen, 1973), coloca os sistemas biológicos em uma espécie de equilíbrio dinâmico. No livro *Alice Através do Espelho* de Lewis Carroll, a Rainha Vermelha diz “é preciso correr o máximo possível para permanecer no mesmo lugar” (“*it takes all the running you can do, to keep in the same place*”). Apesar de parecer loucura, essa é a forma como os sistemas biológicos têm se mantido estáveis ao longo dos anos. Quando duas espécies interagem na natureza, diversos fatores contribuem para que ocorram mudanças evolutivas acopladas. Mudanças em uma espécie geram pressões seletivas que levam a mudanças na outra espécie, e vice-versa, resultando na manutenção da afinidade. Dessa forma, é possível que uma interação se mantenha ao longo de milhares de anos.

O conceito de coevolução, formulado por Ehrlich e Raven na década de 1960, tem sido objeto de discussões ao longo dos anos (Ehrlich and Raven, 1964). De forma simplificada, a coevolução é um processo determinado por mudanças evolutivas recíprocas entre espécies que interagem, guiadas pela seleção natural. Esse processo, entretanto, não pode ser observado, pois ocorre de forma lenta ao longo de muitos e muitos anos. Em seu trabalho, Ehrlich e Raven se perguntam “o que é possível aprender sobre a coevolução de organismos que interagem ecologicamente, sem ter acesso à sua história evolutiva?” (Ehrlich and Raven, 1964). De fato, não há como se ter acesso à história evolutiva das espécies, entretanto, esse processo deixa marcas nos sistemas e, a partir dessas marcas, a história pode ser reconstruída.

O processo de coevolução pode ocorrer, não apenas em sistemas macroscópicos, definidos por interações do tipo presa-predador ou parasita-hospedeiro, mas também em sistemas microscópicos, definidos por interações entre macromoléculas. Da mesma forma que um parasita sofre mudanças em suas características morfológicas, de forma a manter a afinidade pelo seu hospedeiro ao longo dos anos, uma molécula biológica sofre mudanças em sua composição, de forma a manter a afinidade por seus parceiros moleculares.

O objetivo geral deste trabalho é investigar, *in silico*, as marcas deixadas por um processo de coevolução molecular em duas famílias de proteínas, de forma a tentar elucidar os princípios que guiaram esse processo. Considerando a hipótese de que aconteceram, de fato, mudanças evolutivas acopladas nas proteínas parceiras, a intenção é tentar reconstruir uma

rede de afinidades que reproduza, da melhor forma possível, o resultado do processo de coevolução ocorrido ao longo de milhares de anos. A partir dessa rede, será possível definir os princípios que determinam a afinidade e seletividade entre proteínas dessas famílias.

Para realizar a investigação proposta é preciso ter informação sobre as características de um número suficiente grande e variado de proteínas pertencentes a cada uma das duas famílias que serão estudadas. Entretanto, não é possível extrair essas informações de bancos de dados de estruturas 3D, pois muitas proteínas ainda não possuem estrutura determinada. Por conta disso, as análises a serem realizadas devem ser feitas considerando simplesmente sequências de aminoácidos. Essa redução de dimensões (de 3 para 1) causa perda de informação, o que torna o problema mais complexo, e faz necessária uma abordagem teórica sofisticada. A abordagem teórica proposta envolve conceitos e princípios das áreas da biologia, física, matemática, estatística e computação e é capaz de propor modelos que podem ser verificados experimentalmente.

Coevolução Molecular

Interações entre proteínas são a base de sua função molecular (Lovell and Robertson, 2010). Nesse contexto, o conhecimento evolutivo de interações do tipo proteína-proteína é importante para a compreensão de como os sistemas biológicos surgiram e persistiram (Lewis et al., 2010). As diversas substituições de aminoácidos que podem ocorrer ao longo do tempo nas interfaces de interação têm o potencial de modificar a especificidade da ligação e são a chave para as mudanças evolutivas que ocorrem em redes de interação entre proteínas (Lovell and Robertson, 2010).

A interação física entre duas proteínas pode determinar a ocorrência de mudanças evolutivas recíprocas, ou seja, coevolução. De fato, quando dois sítios estão em contato direto, uma mutação no primeiro sítio pode modificar as pressões seletivas que agem sobre o segundo sítio e vice-versa. Às vezes, não apenas o *fitness* de um sítio acoplado é modificado, mas a paisagem de *fitness* da proteína inteira (Lovell and Robertson, 2010).

Pode-se identificar coevolução molecular de diversas maneiras como, por exemplo, através da detecção de mutações correlacionadas em sítios específicos (Madaoui and Guerois, 2008; Mintseris and Weng, 2005). A coevolução pode ser intra ou intermolecular. A maioria do sítios na estrutura de uma proteína apresentam determinado grau de coevolução, visto que contribuem para a integridade estrutural e, de certa forma, para a manutenção da função. Entretanto, alguns sítios estão mais diretamente acoplados do que outros (Lovell and

Robertson, 2010).

A identificação de sítios diretamente acoplados pode contribuir para o estudo de interações do tipo proteína-proteína. Em um trabalho recente de Champeimont et al., uma análise de coevolução molecular foi utilizada como ferramenta para a reconstrução da rede de interações proteína-proteína do genoma do vírus da hepatite C (Champeimont et al., 2016). Em seu trabalho, Champeimont et al. não apenas foram capazes de recuperar informações sobre quais proteínas estariam interagindo, mas também sobre quais resíduos estariam participando da interação.

No presente trabalho, o processo de coevolução molecular que parece ter ocorrido entre neurotoxinas de escorpião e seus alvos moleculares, os canais iônicos de sódio dependentes de voltagem (canais Na_V), será investigado em nível molecular, tendo em vista a obtenção de pistas que ajudem a solucionar questões de afinidade e especificidade, bem como determinar as interações de aminoácidos responsáveis pela ligação das neurotoxinas em seus respectivos sítios.

Canais Iônicos Dependentes de Voltagem

Canais iônicos dependentes de voltagem (VGIC, do inglês *voltage-gated ion channels*) são macromoléculas que permitem o fluxo seletivo de íons através de membranas celulares em resposta a estímulos elétricos (Stock et al., 2013). Embora não se conheça claramente o contexto em que essas estruturas surgiram, acredita-se que os canais iônicos apareceram juntamente com as primeiras formas de vida e que hoje desempenham um papel fundamental na excitabilidade das células (Hille, 2001). Por conta de sua importância em muitos processos celulares e na transdução de sinais, os VGIC são alvos moleculares de uma grande variedade de toxinas (Catterall et al., 2007), incluindo as que alteram as cinéticas de ativação e inativação conhecidas em inglês por *gating modifier toxins*, as quais serão objeto de investigação do presente trabalho.

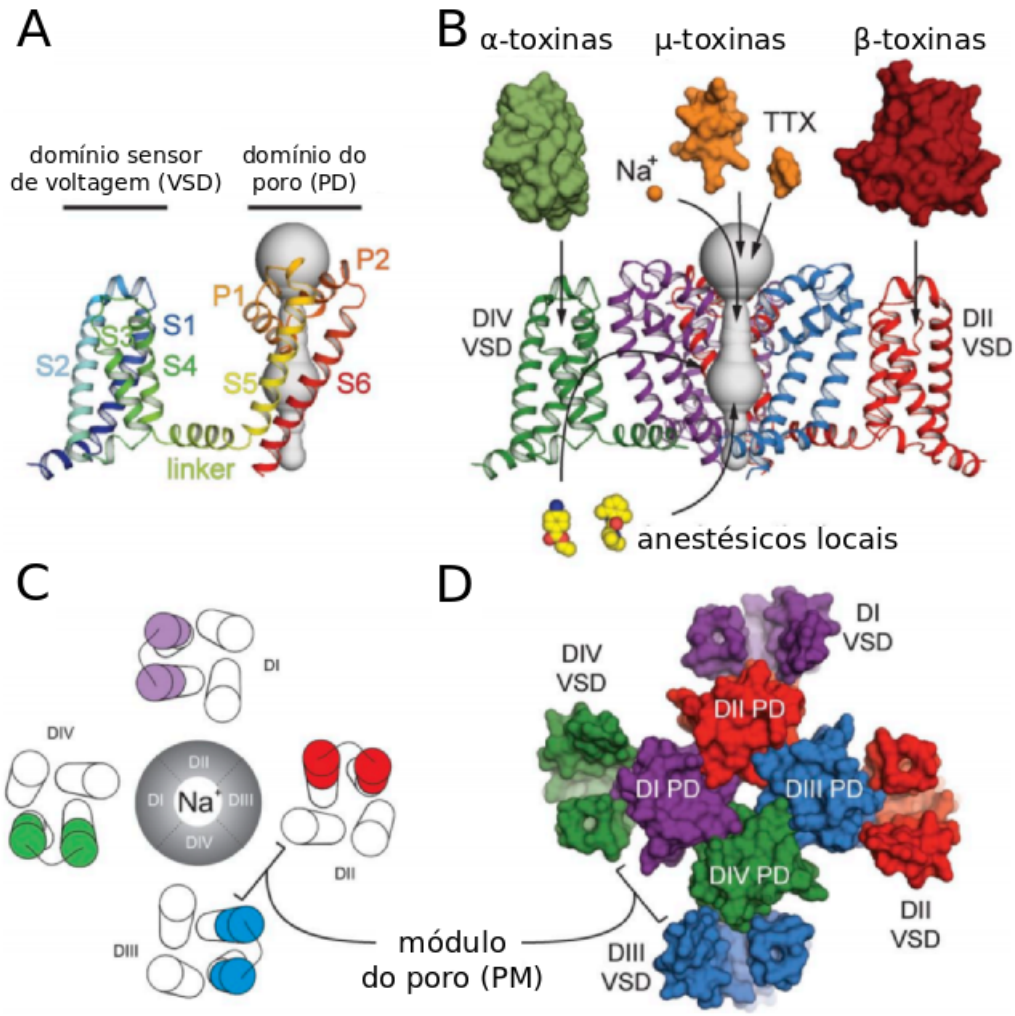

Figura 1: Figura esquemática de canal de sódio dependente de voltagem (Nav). A) Vista lateral de uma subunidade do Nav, composta pelo domínio sensor de voltagem (VSD; hélices S1-S4) e o domínio do poro (PD; hélices S5 e S6). B) Principais ligantes e seus respectivos sítios de ligação em canais Nav. C e D) Vista superios de Nav, mostrando os domínios DI-DIV organizados ao redor do poro central. Adaptado de Ahern et al., 2016.

Os VGIC são codificados por 60 genes no genoma humano e formam uma superfamília com mais de 140 tipos de canais (Yu and Catterall, 2004) que, apesar da grande diversidade, tem aspectos funcionais e estruturais comuns. Os VGIC são formados por quatro subunidades alfa que circundam um poro hidratado central, o qual permite a passagem de íons, e algumas subunidades beta auxiliares. Cada subunidade alfa é composta por seis hélices transmembrânicas, S1-S6 (Figura 1). Os segmentos iniciais, S1-S4, formam o domínio sensor de voltagem (VSD, do inglês *voltage sensor domain*), enquanto o domínio do poro é formado pelos segmentos S5 e S6 (Vargas et al., 2012). Um segmento de ligação (*linker* L) conecta os segmentos S5 e S6 e atua como uma alavanca que regula a mecânica do processo que

promove a abertura e o fechamento do poro iônico (Stock et al., 2013).

O VSD, presente em todas as superfamílias de canais iônicos dependentes de voltagem, é capaz de modificar de forma específica sua conformação em resposta a variações de potencial transmembrânico (Vargas et al., 2012). Quando os primeiros VGIC foram clonados, o quarto putativo segmento transmembrânico, denominado S4, se destacou por conter vários motivos de cargas positivas (argininas e lisinas) (Swartz, 2008). Hoje, sabe-se que resíduos carregados positivamente funcionam como *gating charges*. No repouso, as *gating charges* do segmento S4 são atraídas na direção intracelular pela força eletrostática do potencial de repouso da membrana. Uma vez iniciada a despolarização, essa força eletrostática enfraquece e o segmento se move em direção ao exterior em um movimento em espiral (Vargas et al., 2012).

Os VSD de diferentes tipo de canais podem apresentar funcionamento distinto: os sensores de voltagem de canais iônicos de sódio dependentes de voltagem (Na_V) são mais rápidos dos que os de canais iônicos de potássio voltagem dependentes (K_V). Atualmente, temos provas experimentais de que a maior velocidade de movimentação dos sensores de voltagem de canais Na_V é resultado da presença de resíduos hidrofílicos de Ser e Thr nos segmentos S2 e S4 dos sensores dos domínios I-III e pela presença das subunidades β que se acoplam aos sensores aumentado ainda mais a velocidade de sua resposta (Lacroix et al., 2013). Entretanto, o sensor de voltagem do domínio IV de canais Na_V e K_V tem cinéticas de ativação igualmente lentas e ambos apresentam resíduos hidrofóbicos.

A ativação dos VGIC é seguida por uma uma rápida inativação. Foi constatado que os segmentos S4 dos domínios III e IV ficam imobilizados em uma posição mais exterior durante o processo de inativação rápida, o que evidencia o acoplamento entre a movimentação desses segmentos e o processo de inativação do canal (Cha et al., 1999). Esses resultados indicam que o movimento do segmento S4 do domínio IV é o sinal que dá início ao processo de inativação rápida em canais de Na^+ pelo fechamento do *inactivation gate* (Catterall, 2000).

Toxinas que Modulam a Função de Canais Iônicos

Os atributos em comum dos VGIC (domínio de poro, domínio sensor de voltagem e mecanismos de ativação por voltagem) serviram como alvo para o surgimento de neurotoxinas que alteram o funcionamento desses canais. Em canais de sódio foram descobertos seis sítios onde diversas toxinas, presentes em diferentes organismos, podem se

ligar (Catterall et al., 2007). As toxinas que modificam a função dos VGIC agem por meio de dois mecanismos: *pore-blocking toxins* inibem o fluxo de íons através do canal por meio de um bloqueio mecânico do poro, enquanto as *gating modifier toxins agem* em sítios extracelulares, incluindo o *loop* extracelular entre os segmentos S3 e S4, e modificam a movimentação do sensor de voltagem, alterando a relação entre variações de voltagem e ativação/inativação dos canais (Kalia et al.).

A peçonha de escorpiões contém diversos tipos de neurotoxinas que podem interagir entre si para modular a função de canais iônicos (Quintero-Hernández et al., 2013). A ação conjunta desses polipeptídeos leva à ativação de canais de sódio e à inibição de canais de potássio, causando um prolongamento do potencial de ação em células neuromusculares. Esses eventos levam a um grande influxo de sódio e à liberação de neurotransmissores, seguido por um bloqueio da excitabilidade celular (Gurevitz et al., 2015). Toxinas ativas em canais de Na^+ dependentes de voltagem (NaScTxs, do inglês *sodium channel scorpion toxins*) são polipeptídeos de 6500–8500 Da que apresentam entre 58 e 76 resíduos de aminoácido e quatro pontes dissulfeto, enquanto toxinas ativas em canais de K^+ são constituídas por 42 resíduos e três pontes dissulfeto. Há também os chamados peptídeos birtoxin-like, compostos por 58 resíduos e três pontes dissulfeto. As birtoxinas possuem sequências parecidas, mas com atividade distinta e agem em canais de K^+ (Gurevitz et al., 2015). Além das toxinas encontradas na peçonha de escorpiões, podemos citar ainda a hanatoxina e outras toxinas conhecidas como *cysteine-knot toxins* (CKT) que se ligam ao *loop* extracelular entre S3-S4 de canais de K^+ e inibem a ativação do canal.

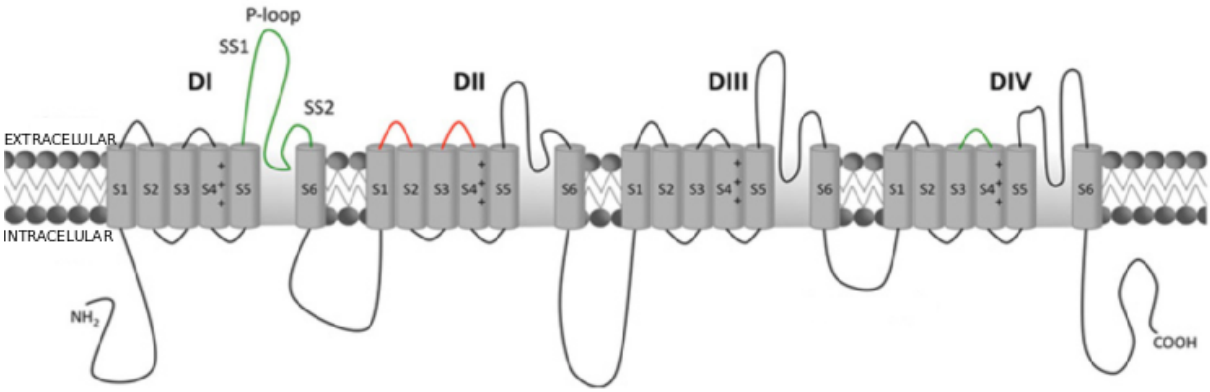

Figura 2: Estrutura de canal de Na^+ dependente de voltagem. As α-toxinas se ligam no sítio 3 (em verde) , enquanto as β-toxinas se ligam ao sítio 4 (em vermelho). Adaptada de Pedraza Escalona and Possani, 2013.

As NaScTxs possuem uma estrutura tridimensional conservada (Figura 3) composta por uma α-hélice e três ou quatro fitas β antiparalelas, que se unem por meio de *loops* irregulares expostos ao solvente e são estabilizadas por quatro pontes dissulfeto espacialmente conservadas (Quintero-Hernández et al., 2013). A hélice se liga à β3 por duas pontes dissulfeto. O par de resíduos de cisteína da hélice é separado por um tripeptídeo, enquanto o par da folha β3 é separado por apenas um resíduo (Pedraza Escalona and Possani, 2013). Numerando os resíduos de cisteína de N- para C-terminal, na alfa hélice estariam os resíduos Cys3 e Cys4 e na folha β3 os resíduos Cys6 e Cys7, formando duas pontes dissulfeto conservadas em todos as NaScTxs descritas: Cys3-Cys6 e Cys4-Cys7. A terceira ponte dissulfeto conservada forma-se entre os resíduos localizados na folha β2, Cys5 e Cys2 (exceto em toxinas anti-inseto excitatórias) (Possani et al., 1999). Essas pontes estão envolvidas na estabilização do núcleo estruturalmente conservado das NaScTxs. A quarta ponte dissulfeto é formada, na maioria das toxinas, entre os resíduos de cisteína N- e C-terminais (Cys1 e Cys8). Nas toxinas anti-inseto excitatórias, a quarta ponte dissulfeto é formada atipicamente entre resíduos contíguos da folha β2 e o resíduos C-terminal (Cys8) (Possani et al., 1999). Existem dois grupos funcionais de NaScTxs: alfa (α) e beta (β), os quais apresentam sítios de ligação distintos em canais Na_V (Couraud et al., 1982).

Sequências primárias dessas toxinas podem ser obtidas por busca específica a partir de sequências de cDNA ou por comparação de similaridade de sequências em análise de transcriptomas de glândulas de escorpiões (Quintero-Hernández et al., 2013). Atualmente, o UniProtKB possui cerca de 800 entradas que correspondem a toxinas de escorpião, sendo mais de 300 entradas correspondentes a NaScTxs.

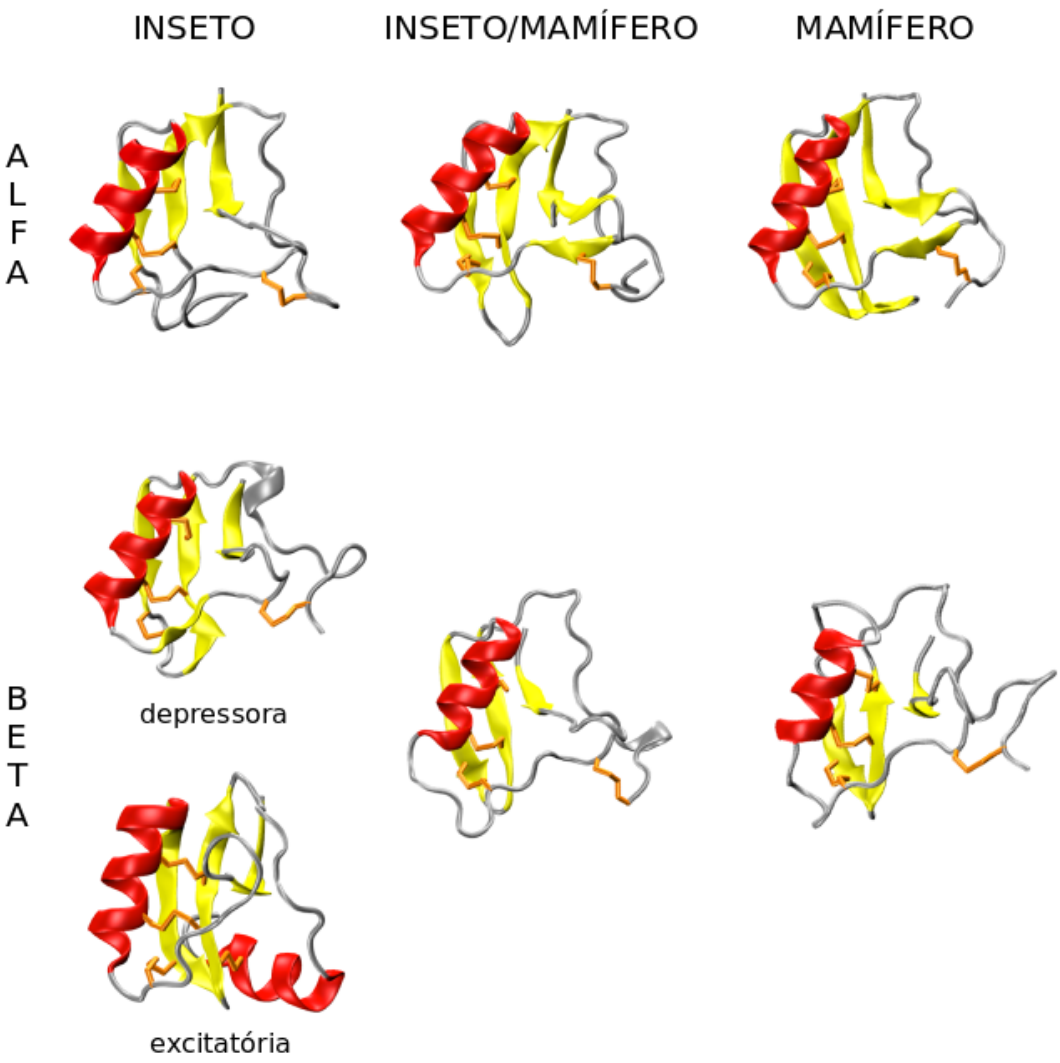

Figura 3: Estrutura tridimensional dos grupos funcionais de NaScTxs. Estão representadas, em sentido horário, BmKαIT1 (PDBID: 2E0H), CsE5 (PDBID: 1NRA), Aah2 (PDBID: 1PTX), Cn2 (PDBID: 1CN2), Ts1 (PDBID: 1B7D), BmKIT1 (PDBID: 1WWN) e LqhIT2 (PDBID: 2I61). As pontes dissulfeto são mostradas em laranja, fitas beta em amarelo e a alfa hélice em vermelho.

Estrutura e Mecanismo de Ação de α-NaScTxs

As chamadas α-toxinas inibem, através da interação com o sítio 3, composto pelo segmento extracelular S3-S4 do domínio IV e pelo segmento SS1-SS2 do domínio I (Figura 2), o processo de inativação rápida em canais Na_V *(Quintero-Hern*ández et al., 2013). A ligação da toxina, que possui grande afinidade pelo estado inativo do canal, impede o

movimento natural do segmento S4 em resposta à despolarização da membrana, levando a um desacoplamento entre os eventos de ativação e inativação rápida do canal (Liu et al., 2005). O resíduo E1613 localizado no *loop* entre os segmentos S3 e S4 do domínio IV está envolvido na ligação das α-toxinas (Quintero-Hernández et al., 2013; Rogers et al., 1996).

Há três subclasses de α-toxinas (Figura 3): 1) α-toxinas clássicas, que agem seletivamente em mamíferos. Alguns exemplos são: Aah2, Aah1 e Aah3 de *Androctonus australis* Hector, Lqq5 de *Leiurus quinquestriatus quinquestriatus* e Bot3 de *Buthus occitanus tunetatus*, 2) α-toxinas anti-inseto, que são altamente tóxicas para insetos. Alguns exemplos são: LqhαIT, Lqq3 e BotIT1 e 3) toxinas α-*like*, que são tóxicas para mamíferos e insetos. Alguns exemplos são: Lqh3 e Lqh6 de *L. quinquestriatus hebraeus*, Bom3 e Bom4 de *B. occitanus mardochei* e BmK M1 de *Buthus martensii* Karsch (Quintero-Hernández et al., 2013).

Foram identificados os seguintes domínios funcionais em α-toxinas: domínio NC (*NC-domain*), domínio do núcleo (*core-domain*) e domínio de ligação (*linker-domain*). O domínio NC é a superfície funcional que contém uma alça de cinco resíduos que se interlaça com o segmento N-terminal (resíduos 8-12) e a região C-terminal (resíduos 56-64). O domínio do núcleo é formado por diversos resíduos (aminoácidos positivamente carregados e hidrofóbicos nos pequenos *loops* que conectam os elementos secundários conservados do núcleo da molécula) espacialmente próximos ao resíduo na posição 18. Os resíduos 8 e 18 dos domínios NC e do núcleo, respectivamente, são interligados pelo domínio de ligação (Quintero-Hernández et al., 2013). A natureza química do domínio do núcleo é altamente conservada dentre as α-toxinas e contém, predominantemente, resíduos de carga positiva e aromáticos/hidrofóbicos (Gordon et al., 2007). O domínio NC varia em composição e arranjo espacial de aminoácidos e provavelmente determina a seletividade da toxina (Gordon and Gurevitz, 2003; Kahn et al., 2009; Karbat et al., 2007; Quintero-Hernández et al., 2013). Foi sugerido que o alto potencial inseticida da toxina LqhαIT esteja associado à conformação protuberante do domínio NC (Figura 4), característica de todas as α-toxinas anti-inseto conhecidas (Karbat et al., 2004). Essa geometria está correlacionada com uma ligação peptídica na conformação *cis* entre os resíduos 9 e 10 da curva de cinco resíduos. Em toxinas anti-mamífero Aah2 e Bmk-M8, a ligação entre os resíduos 9 e 10 está na conformação *trans* (Gordon et al., 2007). Estudos computacionais recentes constataram a formação de diversas pontes salinas e *clusters* hidrofóbicos entre os domínios funcionais de duas α-toxinas (Aah2 e

LqhαIT) e o canal $Na_V1.2$, as quais estabilizam o complexo toxina-VSD (Chen and Chung, 2012; Quintero-Hernández et al., 2013). Os resultados indicam que a superfície funcional das α-toxinas anti-mamífero é centrada no domínio de ligação, de forma semelhante às β-toxinas (Chen and Chung, 2012; Quintero-Hernández et al., 2013).

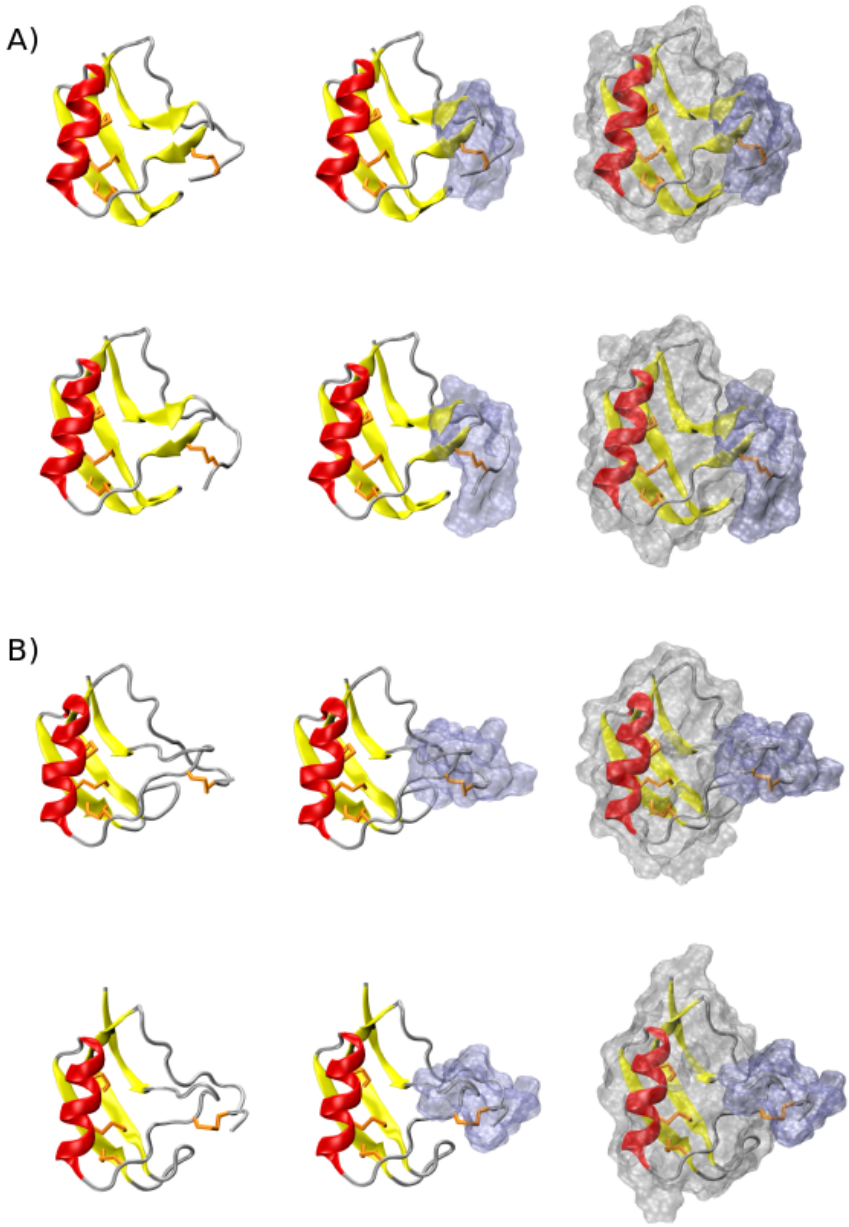

Figura 4: Diferenças entre domínios NC (em azul) de α-toxinas A) anti-mamífero BmK-M8 (acima, PDBID: 1SNB), Aah2 (abaixo, PDBID: 1PTX) e B) anti-inseto Lqq3 (acima, PDBID: 1LQQ), LqhαIT (abaixo, PDBID: 1LQH). As pontes dissulfeto são mostradas em laranja, fitas beta em amarelo e a alfa hélice em vermelho.

Estrutura e Mecanismo de Ação de β-NaScTxs

As β-toxinas se ligam ao sítio 4 de canais Na_V, composto pelos segmentos extracelulares S1-S2 e S3-S4 do domínio II (Figura 2), e modificam o limiar de ativação do canal para potenciais de membrana mais negativos, causando a superativação do canal através de um mecanismo conhecido como *voltage sensor trapping*. Dado um prepulso depolarizante suficientemente forte, o segmento S4 do domínio II move-se para fora, e então a toxina pode se ligar aos resíduos que se tornam acessíveis no *loop* IIS3–S4 e na porção extracelular do segmento IIS4, estabilizando o segmento na posição ativada, e intensificando, assim, a ativação do canal (Cestèle et al., 1998). Há três resíduos de aminoácidos no *loop* extracelular SS2-S6 do domínio III de canais Na_V (Glu837, Leu840, Gly845) que contribuem para a ligação e eficiência da toxina, enquanto os resíduos Asn842, Val843, Glu844, Arg850 e Arg853 atrapalham a ligação (Pedraza Escalona and Possani, 2013) (Figura 5).

Figura 5: Alinhamento de sequências do loop S3-S4 do domínio II de canais Na_V. *Resíduos envolvidos na ligação de β-toxinas, cuja substituição inibe o mecanismo de voltage sensor trapping estão sublinhados (Glu837, Leu840, Gly845). Resíduos cuja substituição aumenta o efeito do voltage sensor trapping aparecem sombreados (Asn842, Val843, Glu844, Arg850 and Arg853), enquanto aqueles cuja subtituição reduz a capacidade da toxina realizar o voltage sensor trapping estão em negrito (Ala841 and Leu846). Retirado de Pedraza Escalona and Possani, 2013.*

Existem quatro subclasses de β-toxinas (Figura 3): 1) β-toxinas anti-mamífero, exclusivas do gênero *Centruroides*, como, por exemplo, os peptídeos Cn2 de *Centruroides noxius* e Css4 de *Centruroides suffusus suffusus*, 2) β-toxinas ativas em insetos e mamíferos, como os peptídeos Ts1 de *T. Serrulatus* e Lqhβ1 de *L. quinquestriatus hebraeus*, 3) β-toxinas anti-inseto seletivas excitatórias, que causam paralisia por contração, como as toxinas AahIT de *Androctonus australis* Hector e Bj-xtrIT de *Hotentota judaica* e 4) β-toxinas anti-inseto

seletivas depressoras, que causam paralisia por flacidez, como o peptídeo LqhIT2 de *L. quinquestriatus hebraeus* (Quintero-Hernández et al., 2013).

Quatro regiões funcionais foram identificadas em β-toxinas: 1) uma região central ("*pharmacopore region*") envolvida na ligação com o sítio do receptor e que consiste em um resíduo negativamente carregado na α-hélice flanqueado por resíduos hidrofóbicos, 2) um *cluster* aromático ("*solvent-exposed aromatic cluster*") que é essencial para a atividade da toxina e está localizado nas folhas β2 e β3, 3) alguns resíduos localizados no centro da *N-groove region*, os quais estão envolvidos no *voltage sensor trapping* e 4) resíduos C-terminais que aumentam a afinidade pelo sítio de ligação no receptor (Cestèle et al., 2006; Cohen et al., 2004, 2005). Diferenças estruturais que explicam a elevada seletividade dessas moléculas estão localizadas nos *loops* variáveis que conectam os elementos principais de estrutura secundária do núcleo e na configuração do *C-tail* (Pedraza Escalona and Possani, 2013).

Hipótese de Trabalho

Este trabalho tem como hipótese principal que seja possível detectar, por meio da análise de sequências primárias de neurotoxinas e canais iônicos dependentes de voltagem, sinais de coevolução molecular que determinem a seletividade e a especificidade entre os pares, assumindo que a coevolução entre tais moléculas tenha, de fato, ocorrido no decorrer de sua história evolutiva. Os estudos realizados serão referentes ao caso particular de coevolução entre domínios sensores de voltagem de canais Na_V e as toxinas de escorpião acima descritas.

Experimento Hipotético e a Busca por um Modelo Otimizado de Coevolução

A principal motivação deste trabalho é a realização de um experimento hipotético que determinaria qual o melhor conjunto de afinidades entre neurotoxinas e canais iônicos, caso todas as barreiras biogeográficas e temporais entre as espécies fossem removidas. Tal conjunto de afinidades definiria um modelo otimizado de coevolução (MOC) para essas moléculas.

O experimento computacional a ser realizado faz analogia a um experimento hipotético que consiste em expor um grande número de neurotoxinas de diferentes tipos ao contato, em ambiente aquoso, com um grande número de canais iônicos de diferentes tipos, inseridos em uma membrana celular infinita por tempo suficiente para que cada neurotoxina

se ligue a algum canal iônico. O esperado é que cada toxina se ligue com maior afinidade a um determinado tipo de canal iônico, apresentando diferenças em termos de afinidade e seletividade.

Objetivos

Objetivo Geral

Investigar a coevolução entre neurotoxinas de escorpião e domínios sensores de voltagem (VSD) de canais Na_V de modo a elucidar os mecanismos moleculares que determinam a especificidade e afinidade entre os pares toxina-VSD.

Objetivos Específicos

1. Analisar um alinhamento múltiplo de sequências de α- e β-toxinas de escorpião, utilizando teoria da informação para identificar os canais de informação que possam explicar a afinidade diferenciada dessas toxinas por sensores de voltagem dos domínios II e IV de canais Na_V;
2. Montar um banco de estruturas tridimensionais de α- e β-toxinas para analisar a superfície dessas moléculas e seu comportamento em água;
3. Reproduzir computacionalmente um experimento hipotético para determinar a seletividade e afinidade entre pares de α- e β-toxinas e domínios sensores de voltagem de canais Na_V;
4. Encontrar um modelo otimizado de coevolução para α- e β-toxinas e domínios sensores de voltagem de canais Na_V;
5. Recuperar as interações moleculares que definem as regras de formação do modelo de coevolução encontrado, de forma a compreender os elementos que caracterizam cada um dos grupos de toxinas.

Metodologia

Quando duas proteínas interagem, contatos físicos entre aminoácidos específicos são estabelecidos. Para determinar quais contatos entre aminoácidos são importantes para a interação de α- e β-toxinas com seus respectivos sítios em canais Na_V, serão construídos dois alinhamentos múltiplos de sequências (*multiple sequence alignment*, MSA): um de α- e β-toxinas, e um de VSD-II e -IV de canais Na_V. Em seguida, serão identificadas posições no

MSA de neurotoxinas que ajudem a diferenciar toxinas do tipo α de toxinas do tipo β. Isso será feito por meio do cálculo da informação mútua (*mutual information*, MI) existente entre o tipo de toxina e o resíduo de aminoácido que aparece em uma determinada posição do MSA. Com essa medida, é possível identificar posições no MSA que ajudem a determinar o tipo da toxina. As posições identificadas dessa forma são candidatas a formar contatos físicos com resíduos dos canais Na_V durante a interação. A análise de MSAs é uma ferramenta que vem sendo utilizada em trabalhos recentes para a identificação de posições de aminoácidos importantes em diferentes famílias de proteínas, sendo frequentemente utilizada em conjunto com análises estatísticas e informacionais (Champeimont et al., 2016; Madaoui and Guerois, 2008; Mintseris and Weng, 2005; Palovcak et al., 2014).

Uma vez definidas as posições de aminoácidos que provavelmente estão interagindo com o canal iônico, será realizada uma simulação computacional para determinar qual é o tipo de canal iônico preferido, em termos de afinidade de ligação, de cada uma das neurotoxinas do MSA. O objetivo é tentar reconstruir uma rede de interações, ou seja, um conjunto de pares toxina-canal, que seja a melhor solução global para o sistema em termos de afinidade e seletividade.

A determinação dessa rede de afinidades, entretanto, é um desafio, pois constitui um problema computacional de ordem n!, que cresce com o número de pares de neurotoxinas e canais considerados no sistema. Como o número de soluções cresce muito rapidamente, torna-se impossível amostrar todo o espaço de soluções. A título de ilustração, 10! possui 7 dígitos, enquanto 100! possui 158 dígitos, e 200!, 375 dígitos. Como solução para esse entrave foi desenvolvido um algoritmo genético (AG) para otimização da rede. O AG inicia sua busca com uma população inicial de redes aleatórias (chamadas indivíduos ou fenótipos) e simula um processo de evolução, de forma que, ao final de um determinado número de gerações, a melhor rede de interações, chamada aqui modelo otimizado de coevolução – MOC, se encontre entre os membros da população.

Após essa etapa, é possível extrair do MOC informações sobre quais posições de aminoácidos na superfície de toxinas e canais Na_V são importantes para manutenção da afinidade entre os pares. Para identificar tais posições, foi, então, realizada uma análise estatística capaz de tornar evidente um conjunto de regras moleculares que definem os tipos funcionais α e β. A partir do conhecimento dessas regras, seria possível, teoricamente, transformar uma α-toxina em β-toxina, e vice-versa.

O trabalho foi executado em duas frentes paralelas, que consistiram em (1) a montagem do sistema que representa um modelo de coevolução entre toxinas de escorpião e VSD de canais Na_V, elemento principal da simulação a ser realizada, (2) a implementação do algoritmo genético que será utilizado para realizar a simulação (Figura 6).

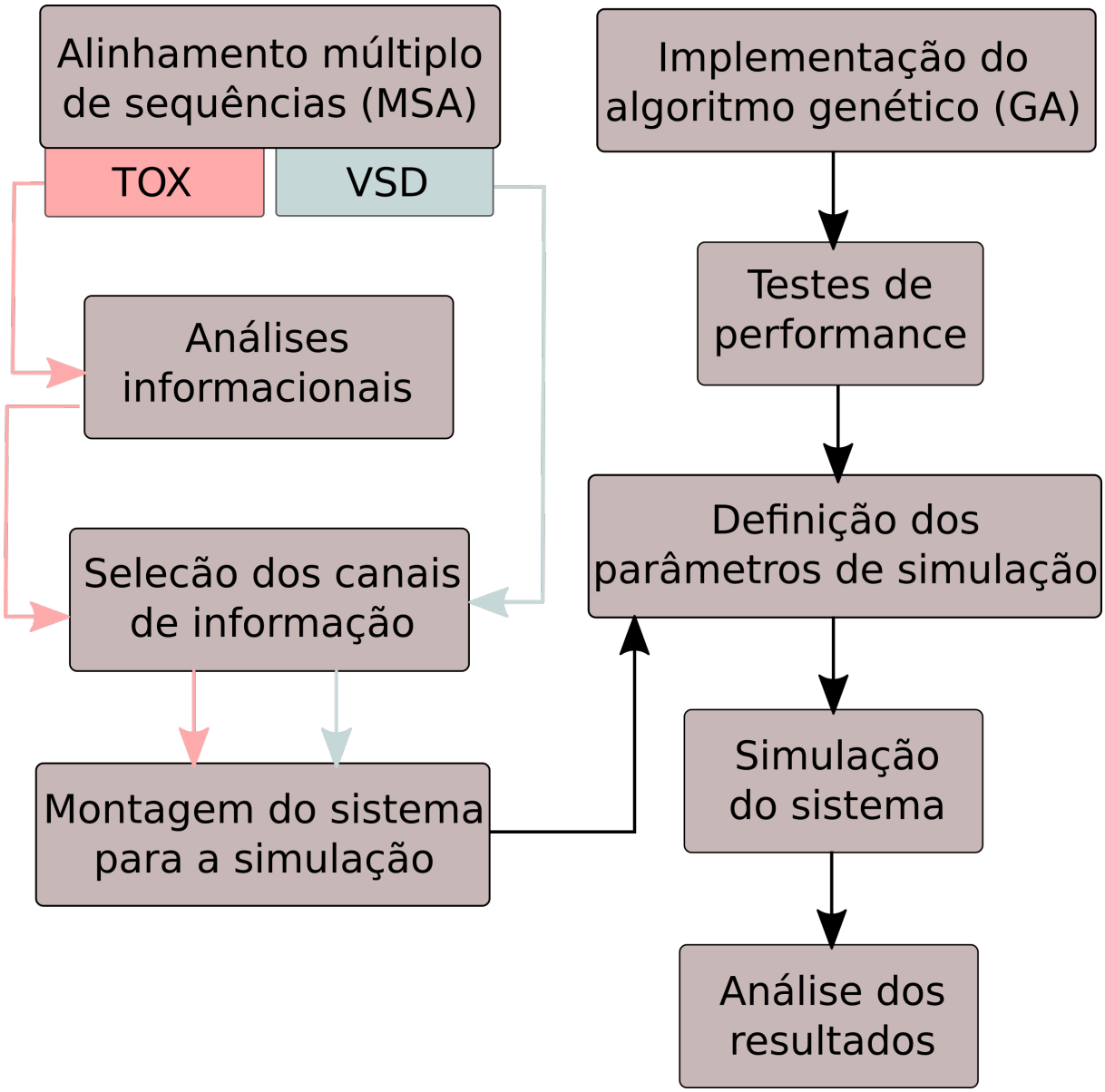

Figura 6: Fluxograma dos principais procedimentos que foram realizados. A sigla TOX é uma abreviação para toxinas e a sigla VSD significa voltage sensor domain.

Alinhamento Múltiplo de Sequências (MSA)

Para montagem de um MSA *seed* de α- e β-toxinas de escorpião, foi realizada uma busca por "*Sodium channel scorpion toxin*" AND ("Alpha" OR "site 3") no UniprotKB, seguida por uma seleção dos *hits* correspondentes a α-toxinas com estrutura tridimensional

depositada no PDB. Analogamente, foi realizada uma busca por "*Sodium channel scorpion toxin*" AND ("Beta" OR "site 4"), seguida por uma seleção dos *hits* correspondentes a β-toxinas com estrutura tridimensional depositada no PDB. Os *hits* selecionados foram, então, alinhados com o programa ClustalX (Larkin et al., 2007).

Uma sobreposição das estruturas das sequências do MSA *seed* foi realizada através da obtenção de uma matriz de rotação a partir da sobreposição dos átomos das cisteínas, seguida pela aplicação da matriz obtida aos demais átomos das proteínas.

Tabela 1: Toxinas do MSA seed, incluindo todas as NaScTxs com estrutura 3D depositada no PDB; a classificação das toxinas foi feita com base em anotações do UniProtKB

Nome	**PDB**	**UniprotKB**	**Classificação**
BmK-M1	1DJT	P45697	*α-like toxin*
Lqq3	1LQQ	P01487	*α-insect toxin*
BmK-M2	1CHZ	P58488	*α-like toxin*
OD1	4HHF	P84646	*α-toxin*
LqhaIT	1LQH	P17728	*α-insect toxin*
BmK-M4	1SN4	P58328	*α-like toxin*
BmK-M10	2KBK	O61705	*α-toxin*
BmK-M7	1KV0	P59854	*α-like toxin*
MeuNaTx-5	2LKB	P86405	*α-toxin*
BmKaIT1	2E0H	Q9GQW3	*α-like toxin*
BmK AGP-SYPU2	2KBH	Q9NJC7	*α-toxin*
BmK-M8	1SNB	P54135	*α-mammal toxin*
AaH2	1PTX	P01484	*α-mammal toxin*
Lqh3	1BMR	P56678	*α-like toxin*
BTN	2A7T	P60277	*α-toxin*
CsE5	1NRA	P46066	*α-toxin*
Kurtoxin	1T1T	P58910	*α-toxin*
CsEv5	1I6G	P58779	*α-toxin*
Cn12	1PE4	P63019	*α-like toxin*
LqhIT2	2I61	Q26292	*β-insect depressant toxin*
Ts1	1B7D	P15226	*β-mammal/insect toxin*
CsEv1	1VNB	P01492	*β-toxin*
	1B3C	Q6V4Z0	*β-toxin*
CsEv3	2SN3	P01494	*β-toxin*

CsEv2	1JZA	P01493	*β-toxin*
CsEI	2B3C	P01491	*β-toxin*
Cn5	2KJA	P45663	*β-toxin*
BmKIT1	1WWN	O61668	*β-insect excitatory toxin*
BmK IT-AP	1T0Z	O77091	*β-insect excitatory toxin*
Css2	2LJM	P08900	*β-mammal toxin*
Cn2	1CN2	P01495	*β-mammal toxin*
Bj-xtrIT	1BCG	P56637	*β-insect excitatory toxin*

Para a montagem de um MSA de NaScTxs, foi realizada uma busca por "*Sodium channel scorpion toxin*" AND ("Alpha" OR "site 3") no UniprotKB, seguida por uma seleção manual dos *hits* correspondentes a prováveis α-toxinas. Analogamente, foi realizada uma busca por "*Sodium channel scorpion toxin*" AND ("Beta" OR "site 4"), seguida por uma seleção manual dos *hits* correspondentes a prováveis β-toxinas. Em seguida, foi gerado um perfil HMM utilizando a função *hmmbuild* do HMMER (Johnson et al., 2010), tendo como base o alinhamento *seed* de toxinas alfa e beta. Por fim, foi realizado o alinhamento dos *hits* (função *hmmalign*) utilizando o perfil gerado no passo anterior.

Para montagem de um MSA de VSDs de canais Na_V, foi realizada uma busca no banco de dados Protein do NCBI por proteínas codificadas pelos genes SCN1A, SCN2A, SCN3A, SCN4A, SCN5A, SCN8A e SCN9A, responsáveis por codificar as subunidades alfa dos $Na_V1.1$ a $Na_V1.7$, respectivamente. Os *hits* com tamanho entre 1500 e 2500 aminoácidos foram selecionados e alinhados com o programa ClustalX (Larkin et al., 2007). Em seguida, foram extraídas desse alinhamento as regiões dos DII-VSD e DIV-VSD. Essas regiões foram alinhadas com o ClustalX e com a função *hmmalign* do HMMER, utilizando o perfil gerado por Palovcak et al. em seu trabalho de 2014 (Palovcak et al., 2014), dando origem a dois alinhamentos distintos da região do *loop* S3-S4.

Foram realizadas e analisadas simulações com os dois tipos de alinhamento da região S3-S4 citados acima. Após análise dos resultados, entretanto, foi constatado que o alinhamento obtido com o HMMER representa melhor a realidade, além de possuir o suporte teórico do trabalho de Palovcak et al., que construíram, testaram e analisaram um alinhamento robusto com mais de 6.000 domínios de VGIC.

Análises Informacionais

Shannon e Weaver afirmam, em seu artigo clássico pubicado em 1949, que o problema fundamental da comunicação é reproduzir, em determinado ponto, de forma exata ou aproximada, uma mensagem emitida em outro ponto. Ignorando-se os aspectos semânticos, é interessante notar que uma determinada mensagem é selecionada de um conjunto de possíveis mensagens. Portanto, um sistema de comunicação deve ser capaz de operar considerando-se cada possível mensagem, e não apenas aquela que realmente será transmitida, pois essa informação é desconhecida à época em que o sistema está sendo projetado (Shannon and Weaver, 1949).

Se o número de mensagens no conjunto é finito, então esse número, ou qualquer função monotônica desse número, podem ser considerados uma medida da informação gerada quando uma mensagem é escolhida, sendo todas as possibilidades igualmente prováveis. Para análises informacionais, a alternativa mais natural é a função logarítmica que, além de ser mais intuitiva, é mais conveniente por motivos matemáticos. A escolha da base logarítmica corresponde à escolha da unidade de medida da informação. Se a base utilizada é a base 2, por exemplo, a unidade resultante pode ser chamada de *binary digits* ou *bits* (Shannon and Weaver, 1949).

No contexto da análise evolutiva de proteínas, pode-se dizer que, dentro de uma determinada família, cada posição da sequência primária representa um canal, por meio do qual informações evolutivas são transmitidas.

A análise de alinhamentos múltiplos de sequências (do inglês *multiple sequence alignment*, MSA) geralmente é feita a partir da extração das frequências sítio-específicas de aminoácidos, ou seja, da distribuição de frequências de cada posição ou coluna do alinhamento. A natureza dessa distribuição é, em geral, informativa sobre a pressão evolutiva que está agindo sobre uma determinada posição (Palovcak et al., 2014). Por exemplo, caso todos os membros de uma família de proteínas globulares apresente uma cisteína conservada em determinada posição, há fortes indícios de que esse resíduo desempenhe importante papel funcional ou de que seja necessário para a manutenção da estrutura globular.

A conservação de um resíduo em determinada posição da sequência está relacionada à entropia daquela posição, ou seja, a quantidade de informação que está sendo transmitida por aquele canal. Uma posição muito conservada apresenta baixos valores de entropia, enquanto uma posição variável apresenta altos valores de entropia. A entropia H de determinado canal i,

em uma família de proteínas, é o somatório negativo das frequências individuais dos resíduos que aparecem naquela posição multiplicadas por sua função logarítmica em base binária, ou seja:

$$H_i = -\sum_A f_i(a) \log_2 f_i(a) \quad (1)$$

, onde A representa o alfabeto de todos os aminoácidos e $f_i(a)$ representa a frequência do aminoácido a no canal i .

A entropia conjunta dos canais i e j é calculada a partir das probabilidades conjuntas dos pares de aminoácidos:

$$H_{ij} = -\sum_{A,B} f_{ij}(a,b) \log_2 f_{ij}(a,b) \quad (2)$$

, onde A,B representa o conjunto de todos os pares de aminoácidos e $f_{ij}(a,b)$ representa a frequência do par de aminoácidos a e b nos canais i e j .

Muita informação também pode ser extraída a partir da distribuição conjunta de frequências, quando considerados pares de posições. No estado nativo, cada resíduo da cadeia polipeptídica realiza interações do tipo resíduo-resíduo. As mutações que ocorrem em determinada posição são, portanto, dependentes da natureza química dos resíduos vizinhos. Como resultado, as distribuições de frequências de aminoácidos em diferentes posições devem ser dependentes umas das outras. A detecção dessas correlações tem o potencial de esclarecer as interações físicas que definem a estrutura nativa (Palovcak et al., 2014). Uma das formas de detectar correlação entre posição é através do cálculo da informação mútua, que pode ser definida como:

$$MI_{ij} = \sum_{a,b} f_{ij}(a,b) \log_2 \frac{f_{ij}(a,b)}{f_i(a) f_j(b)} \quad (3)$$

A entropia condicional, ou a entropia de X dado Y, pode ser calculada como H(X|Y) = H(X,Y) – H(Y) (Figura 7).

A etapa inicial deste trabalho consiste na aplicação de princípios da Teoria da Informação (TI) para detecção de sítios funcionais em NaScTxs, tendo como base um alinhamento robusto de sequências primárias dessas toxinas. Primeiramente, a entropia de cada canal do MSA foi calculada, segundo a fórmula acima descrita, para a identificação de resíduos conservados tanto em α- quando em β-toxinas. Em seguida, foi calculada a informação mútua entre o tipo de toxina e o tipo de resíduo em cada canal. Essa medida foi

obtida de forma indireta, através do cálculo da entropia do tipo de toxina menos a entropia do tipo de toxina condicionada ao tipo de resíduo, segundo a fórmula:

$$MI = H(T) - H(T|A) = -\sum_{T} f(t)\log_2 f(t) + \sum_{T,A} f(a,t)\log_2 f(a,t) - \sum_{A} f(a)\log_2 f(a)$$

, onde $H(T)$ representa a entropia do tipo de toxina, sendo $T = \{\alpha, \beta\}$, e $f(t)$ a frequência do tipo $t \in T$, $H(T,A)$ representa a entropia do tipo de toxina condicionada ao tipo de aminoácido, sendo A o conjunto dos aminoácidos, e $f(a,t)$ a frequência conjunta do tipo t e do aminoácido a . Essa análise foi feita para que fossem detectados sítios determinantes para a diferenciação dos dois grupos de toxinas.

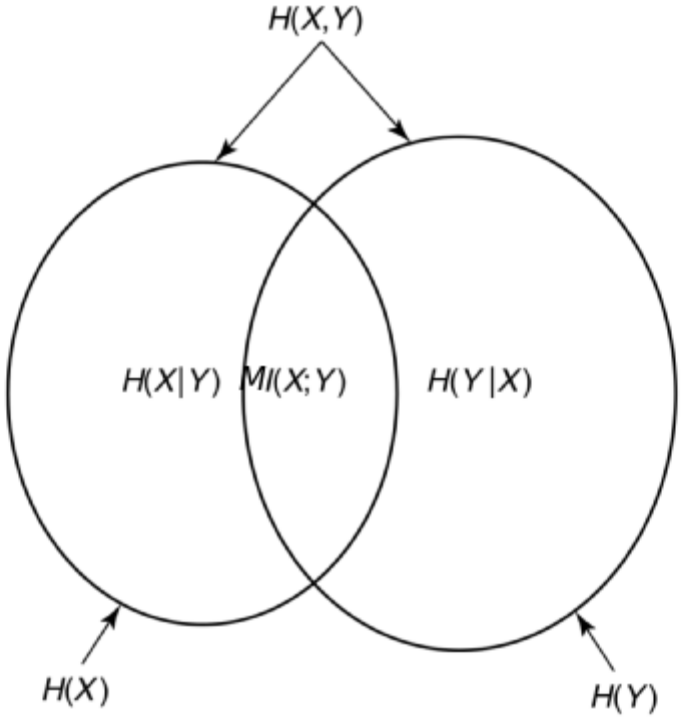

Figura 7: Relação entre entropia (H) e informação mútua (MI). H(X) é a entropia da variável X, H(X,Y) é a entropia conjunta das variáveis X e Y, H(X|Y) é a entropia da variável X dado Y e I(X;Y) é a informação mútua entre as variáveis X e Y. Adaptado de Cover and Thomas, 2012

Tratamento Estatístico dos Dados

Para as análises informacionais do MSA, foi utilizado um alfabeto de cinco letras: 0 – apolar (A, V, I, L, M, F, W, Y, P, C, G); 1 – polar positivo (R, H, K); 2 – polar negativo (D, E); 3 – polar neutro (S, T, N, Q); 4 – gap, conforme a metodologia utilizada por Madaoui e Guerois em seu trabalho (Madaoui and Guerois, 2008). A utilização de um alfabeto reduzido

tem como objetivo melhorar a estatística dos dados.

Além disso, um tratamento estatístico baseado na metodologia utilizada por Morcos et al. (2011) foi utilizado. As frequências utilizadas para os cálculos da entropia, entropia condicional e informação mútua foram determinadas segundo as seguintes fórmulas:

$$f_i(A) = \frac{1}{Meff + \lambda}\left(\frac{\lambda}{q} + \sum_{a=1}^{M}\left(\frac{1}{m^a}\delta_{A,A_i^a}\right)\right) \quad (4)$$

$$f_{ij}(A,B) = \frac{1}{Meff + \lambda}\left(\frac{\lambda}{q} + \sum_{a=1}^{M}\left(\frac{1}{m^a}\delta_{A,A_i^a}\delta_{B,A_j^a}\right)\right) \quad (5)$$

, onde o termo $\delta_{A,B}$ tem valor 1 quando A = B e 0, caso contrário, q representa o número de letras no alfabeto utilizado e λ é um pseudocontador. O termo $\frac{1}{m^a}$ corresponde ao peso de uma sequência e foi inserido para corrigir o viés da amostragem, sendo calculado da seguinte forma:

$$m^a = |\{b \in \{1,\ldots,M\} | seqid(A^a, A^b) > 0.95\}|$$

Ou seja, o termo m^a corresponde à quantidade de sequências que possuem 95% ou mais de similaridade com a sequência a. Por fim, temos o número efetivo de sequências $Meff = \sum_{a=1}^{M}\left(\frac{1}{m^a}\right)$ que representa a soma dos pesos de todas as sequências.

A pseudocontagem é utilizada aqui para garantir que o tamanho limitado do alinhamento de sequências não influencie negativamente os resultados. Isso é feito assumindo-se que as variáveis que não aparecem em determinado canal de informação possuem uma frequência residual. Ou seja, elas poderiam ocorrer, porém não foram amostradas.

Testes de Parametrização

Foram testados alguns parâmetros referentes ao tratamento estatístico dos dados (Tabela 2). Os resultados dos testes indicam que valores crescentes do parâmetro de pseudocontagem λ levam a um aumento dos valores de entropia e de entropia condicional, porém levam à diminuição dos valores de informação mútua (Figura 8).

Mudanças no parâmetro θ levaram a mudanças não direcionais nos valores de entropia

(H), entropia condicional (H|) e informação mútua (MI), ou seja, houve diminuição dos valores para algumas posições e ao aumento para outras.

Todas as análises informacionais realizadas neste trabalho foram feitas utilizando-se os parâmetros $\lambda = 0,5$ e $\theta = 0,95$ (teste 1.3).

Foram calculados ainda alguns valores teóricos para serem utilizados como referência (Tabela 3). Os testes foram realizados com os parâmetros $\lambda = 0,5$ e $\theta = 0,95$. Os casos de referência utilizados foram um canal uniforme, no qual apenas um tipo de aminoácido apareça; um canal, no qual todas as α-toxinas apresentassem resíduos do tipo A e todas as β-toxinas apresentassem resíduos do tipo B; e um canal, no qual resíduos de todos os tipos se apresentassem igualmente distribuídos dentro do grupo das α-toxinas e também dentro do grupos das β-toxinas.

Tabela 2: Parâmetros testados para a análise informacional do alinhamento múltiplo de sequências de α- e β-toxinas de escorpião; θ é o cutoff de similaridade, λ é um pseudocontador e Meff é o número efetivo de sequências (função de theta)

Teste	θ	λ	Meff
1.1	0,95	0,0	162,04
1.2	0,95	0,2	162,04
1.3	0,95	0,5	162,04
2.1	0,99	0,0	227,96
2.2	0,99	0,2	227,96
2.3	0,99	0,5	227,96

*Tabela 3: Valores teóricos de entropia (H), entropia condicional(H|) e informação mútua (MI), com parâmetros estatísticos λ = 0,5 e θ = 0,95; o canal uniforme apresenta apenas um tipo de aminoácidos; o canal 45% Aa, 55% Bb apresenta 45% dos aminoácidos do tipo A (associados às α-toxinas) e 55% dos aminoácidos do tipo B (associados às β-toxinas); o canal 5*9%Xa, 5*11%Xb apresenta todos os tipos de aminoácidos, cada um com frequência 0,09 dentre as α-toxinas e frequência 0,11 dentre as β-toxinas.*

Canal	H (bits)	H| (bits)	MI (bits)
Uniforme	0,029226	0,029212	0,000014
45% Aa, 55% Bb	1,014415	0,029212	0,985203
5*9%Xa, 5*11%Xb	2,313931	2,299132	0,014799

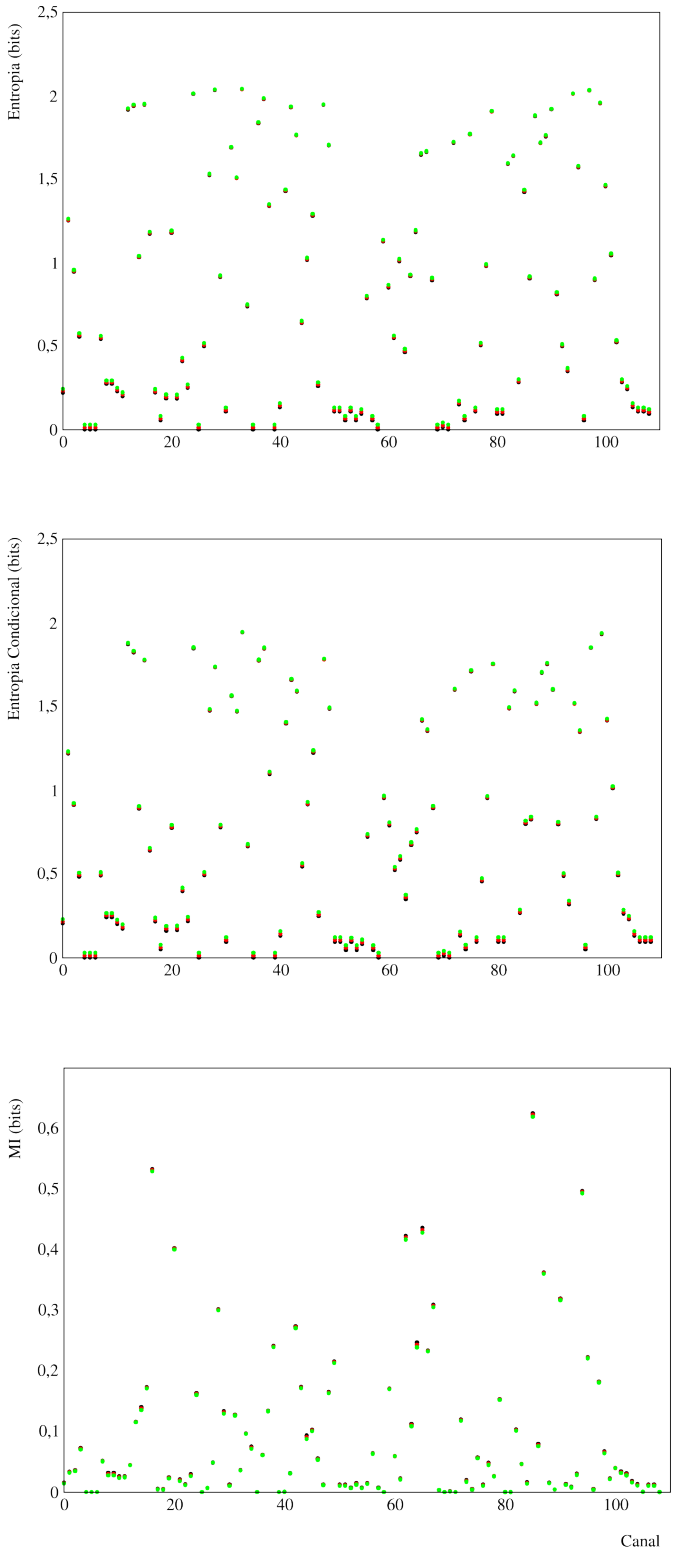

Figura 8: Resultados dos testes realizados para determinação do valor do parâmetro de pseudocontagem (λ) a ser utilizado nas análises informacionais do alinhamento. Em cima, a influência de λ na entropia; no meio, a influência de λ na entropia condicional; abaixo, a influência de λ na informação mútua (MI) calculados para as 109 posições (eixo x) de um alinhamento de α- e β-toxinas de escorpião (n = 242). λ = 0.5, em verde, λ = 0.2, em vermelho e λ = 0.0, em preto.

Seleção dos canais de informação

A seleção dos canais de informação das toxinas foi feita com base na análise informacional do MSA de toxinas. O objetivo da análise é determinar posições que diferem entre os grupos de α- e β-toxinas, mas são conservadas dentro de um mesmo grupos, pois essas são as posições que provavelmente determinam as afinidades diferenciadas de α- e β-toxinas por VSD nos domínios II e IV de canais Na_V.

Já a seleção dos canais de informação dos VSD foi realizada com base em descrições da literatura. Sabe-se que a região mais importante para a interação desse tipo de neurotoxina com o VSD é a região do *loop* S3-S4 e, além disso, sabe-se que, devido à movimentação das hélices S1-S4 durante os eventos de ativação e desativação do canal, alguns resíduos enterrados na membrana podem ser expostos. Diante disso, foram determinados como canais de informação dos VSD, todas as posições do alinhamento pertencentes ao loop S3-S4 mais as 7 posições mais exteriores da hélice S3 e as 7 posições mais exteriores da hélice S4.

Para testar se os canais de informação selecionados para as toxinas são realmente os melhores canais de informação, foram realizados testes em que os canais sofriam variações. Foram definidos quatro tipos de variação: (1) apenas um dos canais era modificado, (2) dois dos canais eram modificados, (3) quatro dos canais eram modificados; e (4) todos os canais eram modificados, ou seja, eram considerados canais aleatórios.

Cada tipo de teste foi realizado 100 vezes. A cada vez, os canais a serem modificados eram escolhidos aleatoriamente dentre os canais de informação originais e substituídos por um canal aleatório dentre os 102 canais restantes. No caso (4), os canais eram simplesmente sorteados dentre as 109 possibilidades. Após a seleção dos canais, era feita uma simulação de 1000 gerações para minimização do sistema.

O sistema utilizado para esses testes consiste em um alinhamento de 205 toxinas, dentre as quais 123 anotadas como β-toxinas e 82 anotadas como α-toxinas, e um alinhamento de 205 VSD, dos quais 121 eram VSD do domínio II e 84 eram VSD do domínio IV. Esses alinhamentos foram obtidos a partir dos alinhamentos originais de toxinas e de VSDs, removendo-se cerca de 97% da redundância nas regiões selecionadas como canais de informação.

Posteriormente, foram gerados 1.000.000 de modelos de coevolução aleatórios para cada um dos 401 conjuntos de canais de informação testados. Dessa forma, foram obtidos 401 espectros de energias, compostos pelas energias dos modelos aleatórios gerados mais a

energia do MOC, que serão mostrados na sessão de Resultados.

Otimização de Modelo de Coevolução

O termo algoritmo genético (AG) refere-se a um algoritmo de busca que simula o processo de seleção natural, utilizando conceitos como hereditariedade, mutação, seleção e recombinação. O objetivo principal é fazer com que uma população inicial de soluções candidatas (chamadas indivíduos ou fenótipos) evolua para soluções melhores. Cada indivíduo possui algumas propriedades (genótipo) que podem sofrer mutação e serem alteradas. A evolução é um processo iterativo, no qual a população de indivíduos vai sendo modificada a cada iteração (ou geração). Em cada geração, o *fitness* (função objetivo da otimização) de cada indivíduo é avaliado e os indivíduos são selecionados para reprodução, sofrendo processos de recombinação e mutação aleatória do genoma e dando origem a uma nova geração de indivíduos. Geralmente, o algoritmo termina a sua execução quando um número determinado de gerações é produzido ou quando um valor satisfatório de *fitness* é alcançado.

Neste trabalho, foi desenvolvido em Python e C++ um AG para a realização de uma versão mais simples do experimento hipotético explicitado na sessão “Experimento Hipotético e a Busca por um Modelo Otimizado de Coevolução”. Essa abordagem foi utilizada por ser impossível amostrar todo o espaço de possíveis soluções, que tem tamanho n!, sendo n o número de possíveis pares toxina-VSD em determinado sistema.

O AG inicia a sua busca a partir de uma quantidade pré-definida de soluções aleatórias, as quais encontram-se sobre determinados pontos de uma paisagem de energia, onde cada ponto representa a energia de uma possível solução. Essas soluções vão, então, sofrendo modificações e movendo-se através dessa paisagem em direção a pontos de menor energia. Caso determinada solução inicial não fique presa em um mínimo de energia local, o esperado é que ela atinja o mínimo global de energia da paisagem. Portanto, quanto mais soluções aleatórias são geradas no início, maior a probabilidade de que alguma delas eventualmente atinja o mínimo global de energia.

Há, no geral, dois componentes que são dependentes do problema em um GA: a codificação dos indivíduos e a função avaliadora (Whitley, 1994). O problema aqui abordado foi codificado da seguinte maneira: um modelo de coevolução, ou seja, um determinado conjunto de afinidades toxina-VSD foi considerado um indivíduo; cada indivíduo possui um

genoma que é uma lista de números inteiros (índices dos VSD), a qual define uma reordenação dos VSD em relação às toxinas, cujas posições são mantidas fixas. Um evento de mutação é definido como a inversão de uma determinada porcentagem de posições no genoma.

A função avaliadora foi definida segundo dois critérios: acoplamento entre os canais e energias de interação (Figura 9). Considerando uma matriz de acoplamento (C) entre os M canais de informação das toxinas e os N canais dos VSD e, para cada par toxina-VSD, uma matriz de energia (E), também da forma MxN, a função avaliadora foi definida da seguinte forma:

$$f = \sum_{P} tr\left(C \cdot E_{a,b}\right)$$

, onde f representa o *fitness* de um indivíduo, C representa a matriz de acoplamento do modelo e $E_{a,b}$ representa a matriz de energia do par (a,b) pertencente ao alfabeto P de todos os pares toxina-VSD do modelo (Figura 9).

Em resumo, o que a função avaliadora faz é ponderar as energias de interação de cada par toxina-VSD, levando em conta o acoplamento entre cada canal de informação. Dessa forma, canais fortemente acoplados que apresentem boas energias de interação devem contribuir substancialmente para a diminuição da energia total do modelo. O esperado, nesse contexto, é que os melhores modelos apresentem as menores energias livres. O algoritmo desenvolvido busca, portanto, minimizar a função de energia acima descrita.

Os parâmetros básicos do algoritmo (taxa de reprodução, taxa de mutação, tamanho da população, passos de minimização, etc) foram testados com um *toy model*, cuja solução de energia mínima poderia ser facilmente inferida. Os resultados obtidos foram então aplicados ao sistema real, com as devidas adaptações de escala. A utilização dos parâmetros adequados é importante, pois influencia no tempo que o algoritmo leva para encontrar a solução de energia mínima e faz com que se torne menos provável a ocorrência de um falso positivo, encontrado em um mínimo local de energia.

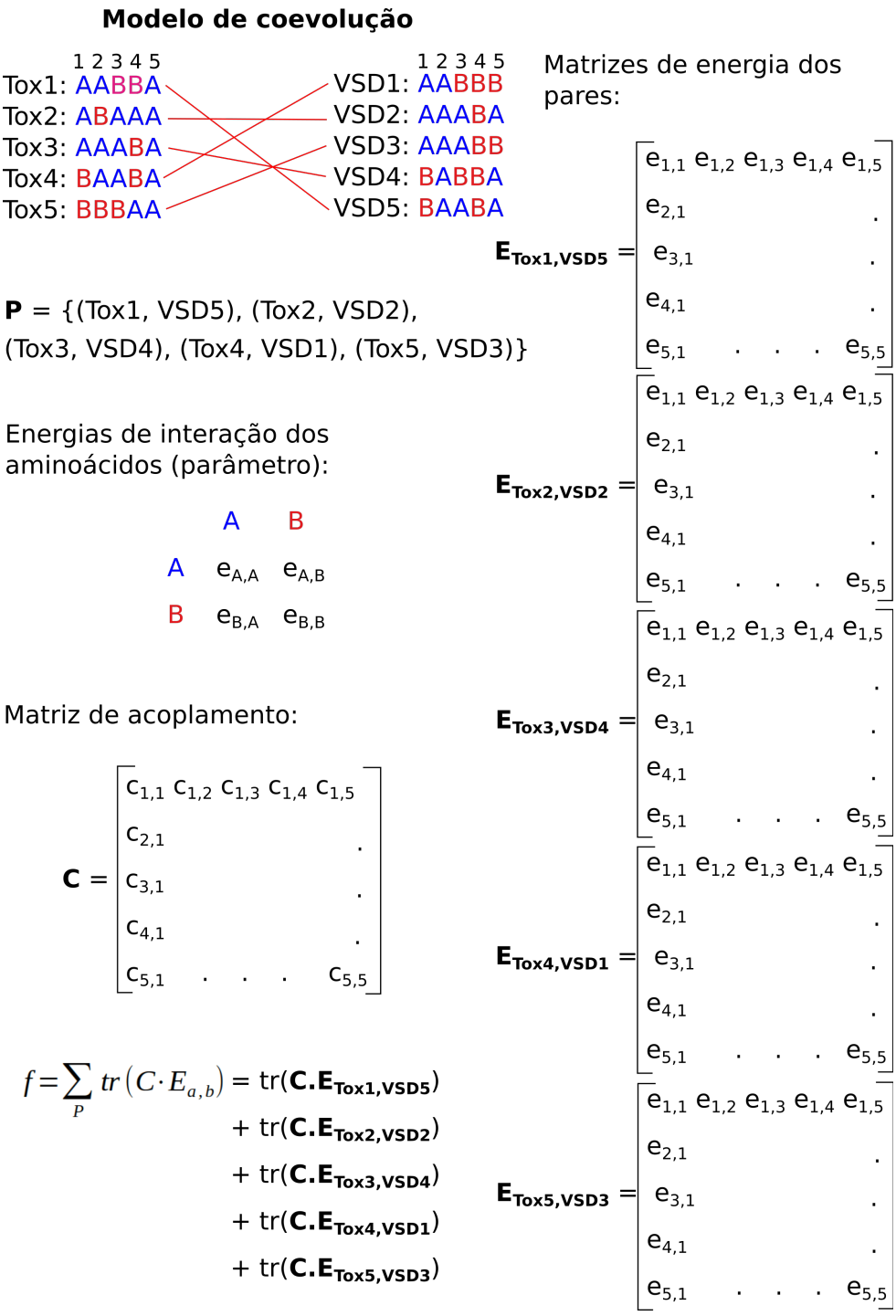

Figura 9: Apresentação dos principais elementos da abordagem teórica desenvolvida. O modelo de coevolução é um conjunto de interações entre pares de NaScTxs e VSD de canais Na_V*, representado por P. A partir das energias de interação entre os diferentes tipos de aminoácidos (definida como parâmetro), são calculadas n matrizes de energia (E), uma para cada par de sequências. A matriz de acoplamento é calculada para o modelo como um todo e representa o acoplamento entre cada par de canais de informação Tox-VSD. A função avaliadora, f, ou fitness, é definida como a soma dos traços do produto entre a matriz de coupling e cada uma das matrizes de energia. O objetivo é chegar ao modelo de energia mínima.*

Testes de parametrização

Foram realizados 18 testes com um sistema reduzido de 42 pares de α- e β-toxinas e VSD-II e -IV de canais Na_V para definir os parâmetros a serem utilizados na minimização do sistema real. Os domínios sensores de voltagem foram recuperados e alinhados com ClustalX. Os parâmetros avaliados nos testes foram a matriz de energia, a utilização de vizinhos estruturais, a utilização de pseudocontador e o número de canais considerados nas toxinas.

Foram consideradas três possíveis matrizes de energia. A matriz E1 (Figura 10) leva em consideração tanto interações que reduzem a energia quanto interações que elevam a energia do modelo, sendo consideradas favoráveis as interações entre resíduos polares de carga neutra e polares de cargas opostas e desfavoráveis as interações entre resíduos polares de mesma carga e interações do tipo polar-apolar. As demais interações não contribuem para a energia do modelo. A matriz E2 (Figura 11) é uma adaptação da matriz E1, onde aparecem apenas as interações que reduzem a energia do modelo. As demais interações são consideradas neutras. A matriz E3 (Figura 12) apresenta as interações que Madaoui e Guerois utilizaram em seu trabalho de 2008, ao definirem o conceito de complementariedade entre os pares aminoácidos nas interfaces de interação proteína-proteína (Madaoui and Guerois, 2008).

O conceito de vizinhos estruturais também vem do artigo de Madaoui e Guerois, 2008. Vizinhos estruturais são resíduos que se encontram espacialmente próximos uns aos outros, podendo estar evolutivamente acoplados, tendo sofrido mutações compensatórias para que alguma complementariedade importante na superfície de interação não fosse prejudicada. Nos testes em que foram considerados vizinhos estruturais, para determinar a energia de interação entre a posição i da toxina A e a posição j do VSD B, foi selecionada a menor dentre as energias de interação da posição i e duas posições vizinhas com a posição j do VSD.

O efeito do pseudocontador nas análises informacionais já foi abordado em uma sessão anterior. No contexto da seleção do melhor modelo de coevolução, a principal função do pseudocontador é a de melhorar a estatística dos dados.

Os canais de informação das toxinas foram selecionados segundo critérios informacionais e são mostrados na sessão de Resultados na Tabela 9. Como canais de informação dos VSDs, foram selecionadas do alinhamento as sete posições que correspondem à parte mais exterior da hélice S3, as 14 posições que correspondem ao *loop* extracelular S3-S4 e as sete posições da parte mais exterior da hélice S4, totalizando 28 canais em todos os testes. Os demais parâmetros fixos dos testes encontram-se na Tabela 4.

Tabela 4: Parâmetros básicos do algoritmo genético utilizados nos testes

Número de gerações	Tamanho da população	Taxa de mutação	Taxa de mortalidade
500	30	10%	20%

	Apolar	Polar +	Polar -	Polar 0	Gap
Apolar	0	1	1	1	0
Polar +	1	1	-1	-1	0
Polar -	1	-1	1	-1	0
Polar 0	1	-1	-1	-1	0
Gap	0	0	0	0	0

Figura 10: Matriz de energia E1. Para determinação das energias de interação entre os pares, os aminoácidos foram agrupados da seguinte forma: apolar – A, V, I, L, M, F, W, Y, P, C, G; polar+ – R, H, K; polar- – D, E; polar0 – S, T, N, Q; e gap.

	Apolar	Polar +	Polar -	Polar 0	Gap
Apolar	0	0	0	0	0
Polar +	0	0	-1	-1	0
Polar -	0	-1	0	-1	0
Polar 0	0	-1	-1	-1	0
Gap	0	0	0	0	0

Figura 11: Matriz de energia E2. Para determinação das energias de interação entre os pares, os aminoácidos foram agrupados da seguinte forma: apolar – A, V, I, L, M, F, W, Y, P, C, G; polar+ – R, H, K; polar- – D, E; polar0 – S, T, N, Q; e gap.

	Apolar	Polar +	Polar -	Polar 0	Gap
Apolar	-1	0	0	0	0
Polar +	0	0	-1	-1	0
Polar -	0	-1	0	-1	0
Polar 0	0	-1	-1	-1	0
Gap	0	0	0	0	0

Figura 12: Matriz de energia E3. Para determinação das energias de interação entre os pares, os aminoácidos foram agrupados da seguinte forma: apolar – A, V, I, L, M, F, W, Y, P, C, G; polar+ – R, H, K; polar- – D, E; polar0 – S, T, N, Q; e gap.

Foi considerado positivo (+) todo teste que conseguiu, ao final de 500 gerações, chegar a um modelo de coevolução que pareasse, com menos de 5% de taxa de erro, toxinas beta com VSD II (23 pares) e toxinas alfa com VSD IV (19 pares). Os testes que não chegaram a tal resultado foram considerados negativos (-).

Tabela 5: Resultados dos testes de parametrização com a matriz de energia E1; foi considerado positivo (+) todo teste que conseguiu, ao final de 500 gerações, chegar a um modelo de coevolução que pareasse, com menos de 5% de taxa de erro, toxinas beta com VSD II (23 pares) e toxinas alfa com VSD IV (19 pares); os testes que não chegaram a tal resultado foram considerados negativos (-)

Teste	Vizinhos estruturais	Pseudo-contador	Número de canais	Resultado	Energia mínima
1.1.1	não	não	12	-	62,44
1.1.2	não	não	7	-	39,18
1.2.1	não	sim	12	-	116,97
1.2.2	não	sim	7	-	71,22
1.3.1	sim	não	12	+	-853,26
1.3.2	sim	não	7	+	-663,02

Tabela 6: Resultado dos testes de parametrização com a matriz de energia E2; foi considerado positivo (+) todo teste que conseguiu, ao final de 500 gerações, chegar a um modelo de coevolução que pareasse, com menos de 5% de taxa de erro, toxinas beta com VSD II (23 pares) e toxinas alfa com VSD IV (19 pares); os testes que não chegaram a tal resultado foram considerados negativos (-)

Teste	Vizinhos estruturais	Pseudo-contador	Número de canais	Resultado	Energia mínima
2.1.1	não	não	12	+	-594,81
2.1.2	não	não	7	+	-383,18
2.2.1	não	sim	12	-	-918,48
2.2.2	não	sim	7	+	-585,07
2.3.1	sim	não	12	+	-1023,93
2.3.2	sim	não	7	+	-730,16

Tabela 7: Resultados dos testes de parametrização com a matriz de energia E3; foi considerado positivo (+) todo teste que conseguiu, ao final de 500 gerações, chegar a um modelo de coevolução que pareasse, com menos de 5% de taxa de erro, toxinas beta com VSD II (23 pares) e toxinas alfa com VSD IV (19 pares); os testes que não chegaram a tal resultado foram considerados negativos (-)

Teste	Vizinhos estruturais	Pseudo-contador	Número de canais	Resultado	Energia mínima
3.1.1	não	não	12	+	-1011,95
3.1.2	não	não	7	+	-711,29
3.2.1	não	sim	12	+	-1056,30
3.2.2	não	sim	7	+	-741,69
3.3.1	sim	não	12	+	-1896,36
3.3.2	sim	não	7	+	-1336,37

Foram realizadas simulações com diferentes tipos de MSA de VSD-II e -IV e de α- e β-toxinas. Isso foi feito para verificar a influência da diversidade de VSDs considerada e da forma como as sequências foram alinhadas nos resultados obtidos, bem como para descartar a influência de fatores externos, como o tamanho do sistema, ou as proporções entre toxinas e VSDs dos diferentes tipos. Os resultados são mostrados na sessão de Resultados.

Todas as simulações foram realizadas com os parâmetros definidos no teste 3.2. Foi decidido utilizar o pseudocontador λ para melhorar a estatística e não considerar vizinhos estruturais, nem interações entre aminoácidos que aumentem a energia para tornar o modelo mais simples. O parâmetro θ foi, primeiramente, aplicado ao alinhamento de toxinas. Para tanto, foi utilizada a função *remove redundancy* do programa Jalview. O parâmetro θ foi aplicado, da mesma maneira, ao alinhamento de VSDs e, em seguida, foram removidas ainda, de forma aleatória, algumas sequências de VSDs para que os dois MSA (de toxinas e VSDs) ficassem com o mesmo número de sequências.

Além disso, foram realizadas três réplicas de uma simulação de 50.000 gerações com o AG para verificar se as redes de afinidades encontradas ao final de cada simulação eram consistentes. O resultado encontrado, comparando-se os quatro MOCs obtidos, foi que cada um apresentou, em média, 2,5% de diferença em relação aos outros e as energias totais obtidas tiveram desvio padrão < 0,1%. Diante disso, foi constatado que o AG teve bom desempenho e eficiência em encontrar uma solução de baixa energia para os sistemas que foram testados. O tempo necessário para a minimização variou em função do número de pares

toxina-VSD, do tamanho da população e do número de canais de informação considerados.

Análise do Modelo Otimizado de Coevolução

O MOC encontrado pelo AG, além de determinar qual o melhor conjunto de pares toxina-VSD, guarda informação sobre quais pares de CIs correlacionados mais contribuem, em média, para baixar a energia total do modelo. Para investigar quais pares de aminoácidos contribuem em conjunto para abaixar a energia total do modelo, foi realizada uma análise de componentes principais (*principal component analysis* – PCA).

A PCA é um procedimento utilizado para tornar evidentes os padrões que existem em um determinado conjunto de dados. A ideia é construir um novo sistema de coordenadas para a representação dos dados. O primeiro eixo desse novo sistema é definido, de forma que as projeções dos dados sobre ele tenham a maior variância possível. Em seguida, um segundo eixo, ortogonal ao primeiro, é definido, de forma a conter a maior variância de projeções dos dados possível, e, da mesma maneira, são definidos os demais eixos do novo sistema de coordenadas. Cada um dos eixos representa uma das componentes principais, que podem ser, então, ordenadas por importância. A análise aqui realizada foi feita com base na matriz de covariância das energias de interação, que será definida a seguir.

Primeiramente, para cada uma das K interações toxina-VSD do MOC, foi definido um vetor de energias $\boldsymbol{e_k}$ contendo as energias de interação entre cada um dos canais de informação das toxinas e cada um dos canais de informação dos VSDs. O vetor $\boldsymbol{e_k}$ tem dimensão n, sendo $n = CI_{TOX} \cdot CI_{VSD}$, onde CI_{TOX} representa o número de canais de informação das toxinas e CI_{VSD} , o número de canais de informação dos VSDs. O traço do vetor $\boldsymbol{e_k}$ define a energia total de um determinado par toxina-VSD. É possível, então, calcular uma matriz de covariância, **Cov**, para esse conjunto de vetores, definida da seguinte maneira:

$$\boldsymbol{Cov} = \langle \boldsymbol{e_k} \cdot \boldsymbol{e_k}^T \rangle$$

, onde k pertence ao conjunto K.

Vale ressaltar que a fórmula utilizada calcula a covariância dos valores absolutos de energia e é uma adaptação da fórmula $\boldsymbol{Cov} = \langle (\boldsymbol{e_k} - \langle \boldsymbol{e_k} \rangle) \cdot (\boldsymbol{e_k} - \langle \boldsymbol{e_k} \rangle)^T \rangle$, com $\langle \boldsymbol{e_k} \rangle$ sendo o vetor nulo. Dessa forma, torna-se mais simples a atribuição de um significado físico aos resultados, visto que, segundo a formulação teórica do problema, não existem interações entre

resíduos que contribuam para o aumento da energia de interação de um par toxina-VSD.

Como **Cov** é uma matriz quadrada, com determinante diferente de zero e, portanto, diagonalizável, ela pode ser escrita em sua forma canônica, ou seja, em termos de seus autovalores e autovetores. Um autovetor **v** de uma transformação linear T é um vetor não nulo, tal que, quando aplicada T sobre **v**, não há mudança na direção, mas apenas no módulo de **v**, que fica multiplicado pelo escalar λ, chamado autovalor. Cada autovalor está associado a um autovetor.

Através do processo de diagonalização, tem-se:

$$\boldsymbol{Cov} = \boldsymbol{R} \cdot diag(\lambda_{1,}\lambda_{2,}\ldots,\lambda_n) \cdot \boldsymbol{R}^{T}$$

, onde R representa uma matriz ortogonal de rotação cuja i-ésima coluna corresponde ao autovetor $\mathbf{v_i}$ de um conjunto de autovetores normalizados de **Cov**, {$\mathbf{v_1}$, $\mathbf{v_2}$, …, $\mathbf{v_n}$}, e $diag(\lambda_{1,}\lambda_{2,}\ldots,\lambda_n)$ é a matriz diagonal dos autovalores λ do conjunto {λ_1, λ_2, …, λ_n}. A partir desse processo, são obtidos os autovetores da transformação R, bem como os autovalores a eles associados.

Cada autovetor obtido no processo acima representa uma componente da PCA, enquanto os autovalores, representam as flutuações quadráticas das projeções sobre os autovetores, ou seja:

$$\sum_{i}^{n} \langle (e_{k,i})^2 \rangle = \sum_{i}^{n} \lambda_i$$

Dessa forma, quanto maior o autovalor associado a uma determinada componente (ou autovetor), maior a contribuição dessa componente para explicar a variância dos dados.

Para analisar os MOCs obtidos com o AG, os vetores lineares que contém as energias de interação para cada par de aminoácidos de um determinado par toxina-VSD foram projetados no subespaço composto pelas componentes que explicam a maior parte da variância das energias. As nuvens de dados resultantes da projeção foram, então, analisadas para verificar se houve algum tipo de separação que explique as afinidades diferenciadas de α- e β-toxinas por VSD-IV e -II respectivamente. As análises foram realizadas utilizando o módulo SciPy.

Construção de um Banco de Estruturas

Para a construção de um banco de estruturas de α- e β-toxinas, foi feita uma análise de similaridade em cima das sequências do MSA de toxinas. Foi realizado o cálculo da distância

de Hamming par a par e, após a obtenção da matriz de distâncias, as toxinas foram agrupadas em *clusters* de similaridade com o método DBSCAN do módulo Scikit-learn (Pedregosa et al., 2011). Em seguida, foi verificada a ocorrência de *templates* (toxinas com estrutura 3D determinada) dentro dos *clusters* e, então, as toxinas que não possuíam estrutura 3D foram modeladas com base em algum *template*, junto ao qual elas estavam agrupadas. As modelagens foram feitas com o *software* MODELLER (Webb and Sali, 2014).

Após a modelagem, foram realizadas simulações de dinâmica molecular (DM) de cada toxina em água, inclusive daquelas que já possuíam estrutura cristalográfica determinada. Para tanto, foi empregado o código de dinâmica molecular NAMD (Phillips et al., 2005). Foi utilizada a dinâmica de Lagevin para a manutenção da temperatura em 300K e pressão em 1 atm. O campo de força utilizado foi o CHARMM36 (Huang and MacKerell, 2013). A fim de minimizar efeitos de borda nos limites do sistema, foram aplicadas condições periódicas de contorno nos eixos x, y e z. Os sistemas tiveram trajetória de equilíbrio calculada em 10ns, após um período de 2,3ns utilizando potenciais harmônicos para restringir o movimento da proteína, de forma que cada simulação teve um período total de aproximadamente 12,3ns. Durante esse período, as simulações foram analisadas qualitativamente com o auxílio do VMD (Humphrey et al., 1996).

Para avaliar a estabilidade estrutural das proteínas do banco de dados obtido, foram feitos cálculos RMSD (do inglês, *root-mean-square deviation*) no tempo para os átomos pesados (átomos da cadeia principal que não sejam hidrogênios) de todas as toxinas, durante os 10 ns após o uso de potenciais harmônicos. Após isso, foi calculada a média dos RMSDs obtidos. Também foi calculada, para todos os aminoácidos em cada toxina, a área da superfície acessível ao solvente (SASA, do inglês s*olvent acessible surface area*) média, que diz o quanto a área de um determinado aminoácido contribuiu, em média, para a área de superfície da proteína ao longo da simulação, ou seja, é a média dos valores de SASA de determinado aminoácido em cada *frame* de simulação.

Com o objetivo de caracterizar a superfície das toxinas do banco de dados gerado, foi calculado o perfil de acessibilidade médio, que consiste na média entre os valores de SASA médios para cada posição do MSA. Essa medida indica o quanto os aminoácidos de cada coluna, em média, estão expostos na superfície das toxinas analisadas.

Resultados

Canais de Informação nas Toxinas

O MSA de toxinas obtido possui um total de 290 sequências, 115 anotadas como α-toxinas e 175 como β-toxinas; com um total de 109 colunas de aminoácidos alinhados. Já o MSA de VSDs de canais Na_v possui 329 sequências não idênticas, sendo 133 sequências do domínio II e 196 sequências do domínio IV; com 95 colunas de aminoácidos alinhados.

Uma análise de entropia, utilizando os parâmetros $\lambda = 0,5$ e $\theta = 0,95$ (teste 1.3), foi empregada para avaliar as posições conservadas no MSA de toxinas. Foi considerada conservada toda posição que apresentasse valores de entropia (H) abaixo de 1,0 *bit*. A Tabela 8 mostra os valores de entropia calculado para cada posição conservada, bem como o aminoácido de maior frequência naquela posição. Foram encontradas 28 posições conservadas.

A análise da informação mútua (MI) entre o tipo de aminoácido e o tipo de toxina revelou posições que são conservadas dentro de um mesmo grupo, α ou β, porém diferentes entre os dois grupos (Tabela 9). Foram selecionadas posições que apresentam valores de MI normalizada pela entropia ($MI_{norm.}$) maiores que 0,10 (10% de ganho informacional) e valores de entropia maior que 1,0 *bit* (posições não conservadas). Além disso, para uma determinada posição ser considerada relevante, deveria apresentar, dentro de um mesmo grupo, predominância de pelo menos 50% de um mesmo tipo de aminoácido. As posições consideradas relevantes estão marcadas em cinza na Tabela 9. As posições que apresentaram 60% ou mais de predominância de um único tipo de aminoácido em um dos grupos foram consideradas muito relevantes e estão marcadas em cinza escuro.

Os resultados dos testes realizados com os canais de informação, descritos na sessão Seleção dos canais de informação são mostrados na Tabela 10 e na Figura 13. Os espectros de energia foram ordenados em ordem crescente de energia do MOC e são mostrados na Figura 14. Em cada um dos espectros de energia mostrados, é possível perceber uma grande densidade de pontos no intervalo que vai de 0 a -500, correspondentes às energias de redes de interação geradas aleatoriamente. Há, então, um *gap* que diminui de tamanho no sentido positivo do eixo x entre os pontos na parte superior e o ponto de energia mínima.

Tabela 8: Posições conservadas (H < 1,0 bit) no alinhamento múltiplo de sequências de α- e β-toxinas de escorpião; em cinza, as oito cisteínas responsáveis pela formação das quatro pontes dissulfetos altamente conservada dentre essa família de toxinas

Posição	**H (*bits*)**	**Aminoácido**
3	0,96	R/K
4	0,58	D/E
5	0,03	G
6	0,03	Y
7	0,03	Apolar
8	0,56	Apolar
20	0,21	C
23	0,43	Apolar
26	0,03	C
27	0,52	Apolar
35	0,75	Y
36	0,03	C
40	0,03	C
48	0,28	Apolar
55	0,08	G
57	0,80	Y
59	0,03	C
61	0,87	Apolar
62	0,56	Apolar
69	0,91	Apolar
70	0,03	C
71	0,04	W/Y
72	0,03	C
77	0,13	L
78	0,52	P
79	0,99	D/E
87	0,92	Apolar
97	0,08	C

Tabela 9: Resultados da análise de informação mútua (MI) entre o tipo de aminoácido e o tipo de toxina; são mostradas as posições que apresentam valores de MI normalizada pela entropia ($MI_{norm.}$) maiores que 0,10 (10% de ganho informacional) e valores de entropia (H) maior que 1,0 bit; H| representa a entropia condicional em relação ao tipo de toxina; foi considerada variável (var.) toda posição que não apresentasse pelo 50% de predominância de um único tipo de aminoácido; em cinza claro: posições com pelo menos 50% de predominância, em cinza escuro: posições com 60% ou mais de predominância

Posição	MI (bits)	H (bits)	H| (bits)	$MI_{norm.}$	Alfa	Beta
15	0,14	1,04	0,91	13,0%	Var.	Gap
17	0,53	1,19	0,66	44,7%	N(86%)	G(93%)
21	0,40	1,19	0,79	33,5%	Apol.(63%)	K(94%)
29	0,30	2,04	1,74	14,7%	Gap	Var.
39	0,23	1,34	1,11	17,8%	L(60%)	E(82%)
43	0,27	1,94	1,66	14,0%	K(53%)	Var.
50	0,21	1,71	1,49	12,5%	S(82%)	Var.
60	0,17	1,14	0,97	14,9%	Var.	Y(52%)
63	0,42	1,02	0,61	40,1%	Var.	Gap
66	0,43	1,20	0,77	35,8%	Var.	Gap
67	0,23	1,66	1,42	14,0%	G(66%)	Var.
68	0,31	1,67	1,36	18,3%	N(50%)	Var.
86	0,62	1,44	0,82	43,1%	Gap	Var.
88	0,36	1,88	1,52	19,1%	K/R(77%)	Var.
91	0,32	1,92	1,60	16,5%	G(78%)	Polar
95	0,49	2,01	1,52	24,5%	Gap	N(50%)
96	0,22	1,58	1,36	14,0%	K(56%)	Var.
Média	0,09	0,86	0,77	8,4%		

Tabela 10: Testes de otimização de um mesmo sistema, composto por 205 toxinas e 205 VSDs, com diferentes conjuntos de canais de informação; o controle corresponde ao sistema com os canais selecionados usando TI, e as linhas seguintes, ao mesmo sistema com conjuntos de canais gerados a partir de 1, 2, 4 e 7 mudanças aleatórias no conjunto original; as energias foram computadas após 5.000 gerações do algoritmo genético

Teste	Energia média	Desvio padrão	n
Controle	-3104,8		1
1 mudança	-2583,3	7,5%	100
2 mudanças	-2294,0	9,9%	100
4 mudanças	-1899,6	14,3%	100
Canais aleatórios	-1417,1	21,7%	100

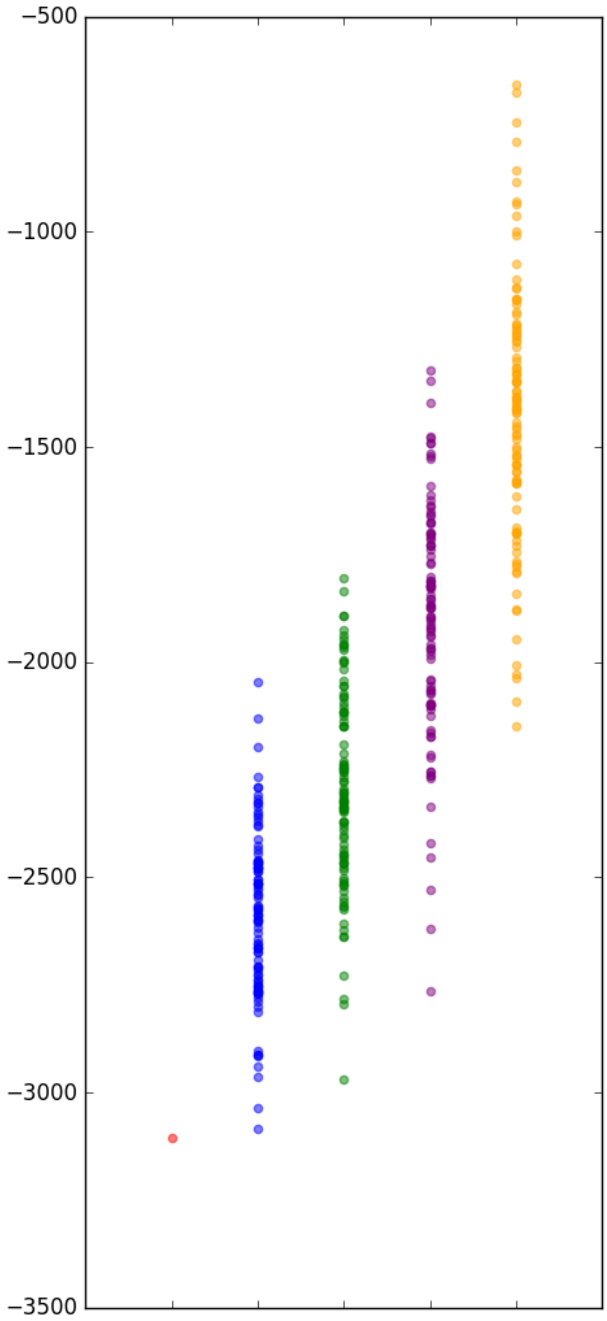

Figura 13: Espectros de energia dos MOCs de um mesmo sistema com diferentes canais de informação nas sequências de toxinas. O ponto vermelho mostra a energia do MOC do sistema com os canais selecionados com TI; os pontos azuis (n = 100), as energias de MOCs do sistema com conjuntos de canais gerados a partir de uma mudança nos canais selecionados; os pontos verdes (n = 100), as energias de MOCs do sistema com canais gerados a partir de duas mudanças nos canais selecionados; os pontos roxos (n = 100), as energias de MOCs do sistema com conjuntos de canais gerados a partir de quatro mudanças nos canais selecionados; e os pontos amarelos (n = 100), a energia de MOCs do sistema com canais aleatórios. Cada ponto foi obtido em uma minimização de 5.000 gerações utilizando o algoritmo genético.

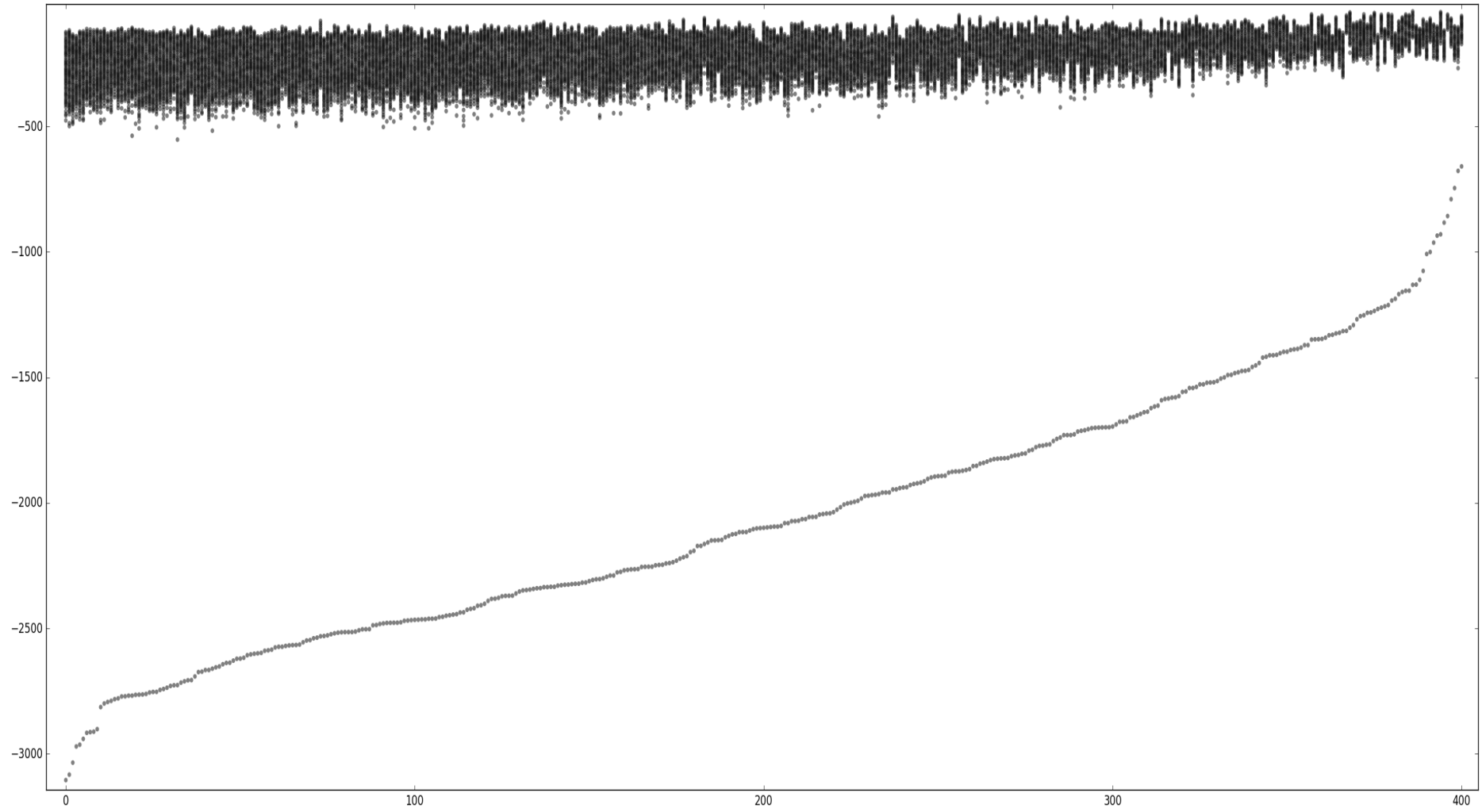

Figura 14: Espectros de energias de modelos de coevolução aleatórios (n = 1.000.000; borrão na região superior) e energia do modelo otimizado de coevolução (pontos na região inferior) para diferentes sistemas. Cada espectro de energias representa o mesmo sistema com diferentes canais de informação nas sequências de toxinas. O primeiro espectro (posição 0 no eixo x) é composto por energias obtidas utilizando os canais de informação selecionados através de análises informacionais do MSA de toxinas. Os espectros estão dispostos em ordem crescente de energia mínima e totalizam 401.

Modelo Otimizado de Coevolução

A Tabela 11 apresenta os resultados de todas as simulações que foram realizadas. Serão apresentados em detalhe os resultados obtidos para os sistemas marcados em cinza. Os sistemas serão chamados sistema 1, sistema 2, sistema 3 e sistema 4, segundo a ordem em que aparecem na tabela. Para todos os sistemas simulados, foi obtido um MOC, no qual α- e β-toxinas formavam dois grupos distintos, interagindo, respectivamente, com VSD-IV e -II. A taxa de erro foi menor do que 5% em todos os casos. Esse resultado é mostrado para o sistema 3 na Figura 15.

Os sistemas 1 e 3 são constituídos por canais $Na_V1.1$-1.7 e toxinas de escorpião de todos os tipos. Já os sistemas 2 e 4 são constituídos por canais $hNa_V1.1$-1.7 e o canal Na_V de *Drosophila melanogaster,* e toxinas que agem em mamíferos ou em insetos e mamíferos. Os CIs considerados nas toxinas são os 12 canais marcados em cinza na Tabela 9. Os CIs dos VSDs totalizam 28 para os sistemas 1 e 2 e são constituídos por posições do *loop* extracelular S3-S4 dos VSD-II e -IV alinhados pelo programa ClustalX. Já nos sistemas 3 e 4, os CIs dos VSDs totalizam 20, sendo constituídos pela mesma região de VSD-II e -IV alinhados conforme alinhamento robusto de mais de 6.000 VSDs publicado no trabalho de Palovcak e colaboradores (Palovcak et al., 2014).

O perfil de aminoácidos dos canais de informação considerados, obtido com a função *hmmbuild* do HMMER, é mostrado, para os sistemas 1, 2, 3 e 4 na Figura 16. Os sistemas 1 e 3 foram construídos de forma a considerar a maior variabilidade possível de sequências de NaScTxs e de canais $Na_V1.1$-1.7. Foi eliminada 96% da redundância entre as sequências de α-toxinas ($\theta = 0,96$) e 97% ($\theta = 0,97$) da redundância entre as sequências de β-toxinas. Também foi eliminada a redundância entre as regiões alinhadas de VSD-II e VSD-IV.

Os sistemas 2 e 4 são redundantes em termos de sequências de VSDs, possuindo 20 cópias de VSD-II de cada tipo de canal Na_V humano ($hNa_V1.1$-1.7) e 11 cópias do VSD-II de *Drosophila melanogaster*, e 38 cópias de VSD-IV $hNa_V1.1/1.2/1.3/1.6$, 40 cópias VSD-IV $hNa_V1.4/1.5$, 20 cópias VSD-IV $hNa_V1.7$ e 9 cópias do VSD-IV de *Drosophila melanogaster*. Em linhas gerais, os sistemas 1 e 3, e 2 e 4 reproduzem, respectivamente, as características de um sistema natural e de um sistema de teste laboratorial.

Tabela 11: Lista dos sistemas que foram simulados; os valores de energia do melhor modelo foram computados após 50.000 gerações; θ é o parâmetro que indica o cutoff de similaridade aplicado às sequências

VSDs		Toxinas		Características	θ				Número de ICs		Energia	Média DII	Média DIV	Média
DII	DIV	β	α		DII	DIV	β	α	TOX	VSD				
123	82	121	84	Na_V1.1-1.7; loop s3-s4	0,99	0,98	0,97	0,96	7	28	-3111,6	-10,4 ± 1,6	-22,4 ± 2,0	-15,2 ± 6,2
									12	28	-4541,5	-16,2 ± 1,8	31,1 ± 3,2	-22,2 ± 7,7
131	111	143	99	hNa_V1.1-1.7 e Na_V inseto; loop s3-s4	0,99	0,99	0,99	0,99	7	28	-4260,4	-12,2 ± 2,7	-24,0 ± 2,0	-17,6 ± 6,4
									12	28	-5632,0	-16,4 ± 1,9	-31,4 ± 2,7	-23,3 ± 7,8
35	20	75	42	Na_V obtidos com perfil HMM; loop s3-s4; toxinas redundantes	0,97	0,94	0,90	0,87	7	28	-	-	-	-
									12	28	-3903,9	-65,3 ± 16,2	-80,9 ± 9,7	-71,0 ± 16,1
131	111	143	99	Na_V obtidos com perfil HMM; loop s3-s4	0,99	0,99	0,99	0,99	7	28	-4141,5	-11,1 ± 2,9	-24,3 ± 2,3	-17,1 ± 7,1
									12	28	-5799,9	-16,6 ± 3,3	-32,7 ± 3,1	-24,0 ± 8,7
105	96	102	99	Na_V1.1-1.9; loop s3-s4	0,99	0,98	0,95	0,97	7	28	-3684,6	11,4 ± 2,0	-25,9 ± 1,9	-18,3 ± 7,5
									12	28	-	-	-	-
131	107	134	104	Tox-mam; hNa_V1.1-1.7; loop s3-s4	-	-	-	-	12	28	-5876,7	-17,1 ± 2,4	-34,0 ± 2,8	-24,7 ± 8,8
123	82	121	84	Na_V1.1-1.7; s3-s4 sem gaps	0,99	0,98	0,97	0,96	12	20	-3489,4	-17,1 ± 2,2	-16,9 ± 1,5	-17,0 ± 2,0
96	103	100	99	Na_V1.1-1.7; s1-s2 e s3-s4 sem gaps	0,99	0,99	0,95	0,99	12	40	-6407,6	-30,8 ± 3,0	-33,5 ± 3,8	-32,2 ± 3,7
105	96	102	99	Na_V1.1-1.9; s3-s4 sem gaps	0,99	0,98	0,95	0,97	12	20	-4662,5	-23,5 ± 3,1	-22,9 ± 2,0	-23,2 ± 2,6
131	107	134	104	Tox-mam; hNa_V 1.1-1.7 e Na_V inseto; s3-s4 sem gaps	-	-	-	-	12	20	-4361,0	-18,4 ± 2,5	-18,2 ± 1,3	-18,3 ± 2,1
131	107	134	104	Tox-mam; hNa_V 1.1-1.7 e Na_V inseto; s1-s2 sem gaps	-	-	-	-	12	20	-5355,3	-21,2 ± 2,5	-24,1 ± 2,9	-22,5 ± 3,1
131	107	134	104	Tox-mam; hNa_V 1.1-1.7 e Na_V inseto; s1-s2 e s3-s4 sem gaps	-	-	-	-	12	40	-8924,1	-36,6 ± 4,2	-38,5 ± 4,3	-37,5 ± 4,4

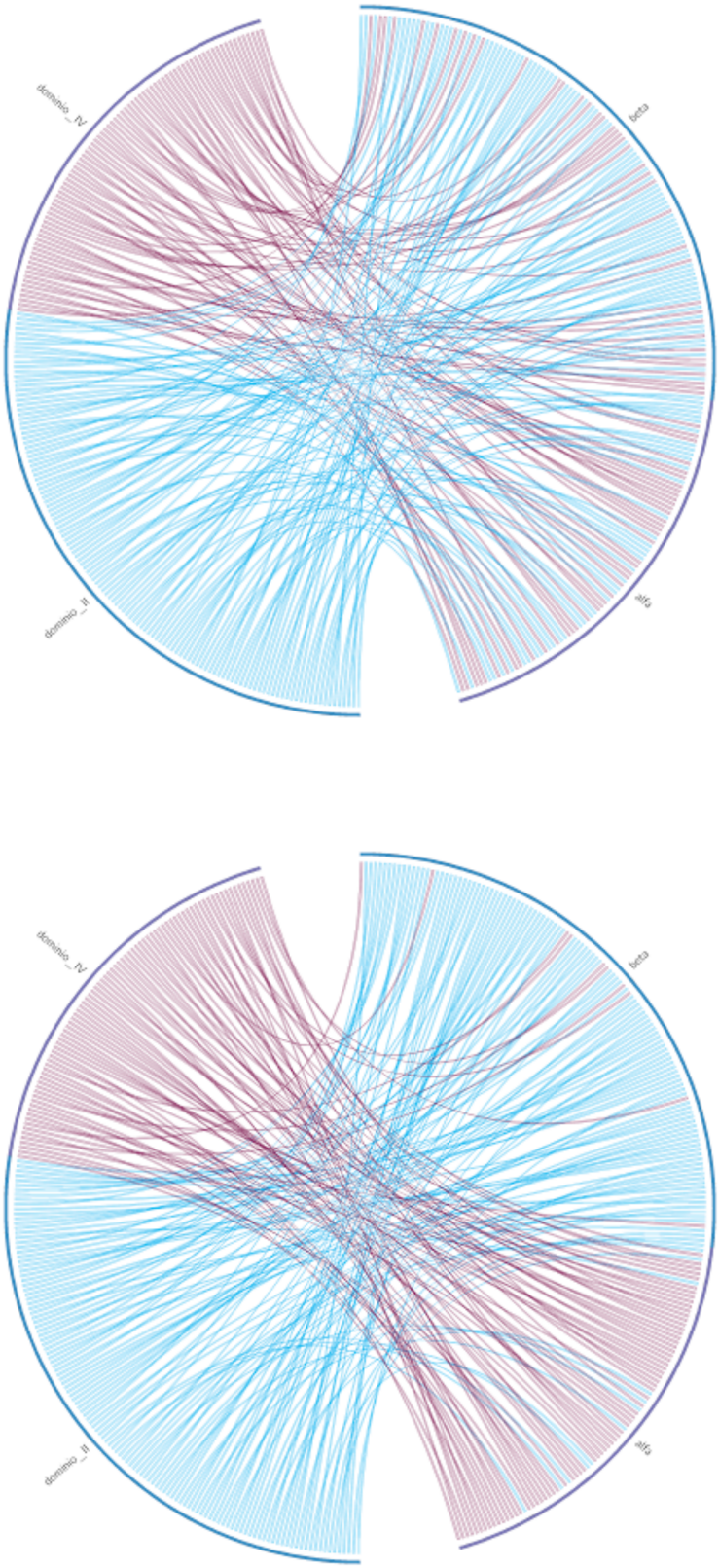

Figura 15: Rede de interações aleatória (acima) e do MOC - sistema 3 (abaixo). As interações com VSD-II são mostradas em vinho e as interações com VSD-IV são mostradas em azul. As β-toxinas estão posicionadas na região superior da metade direita do círculo e as α-toxinas na região inferior.

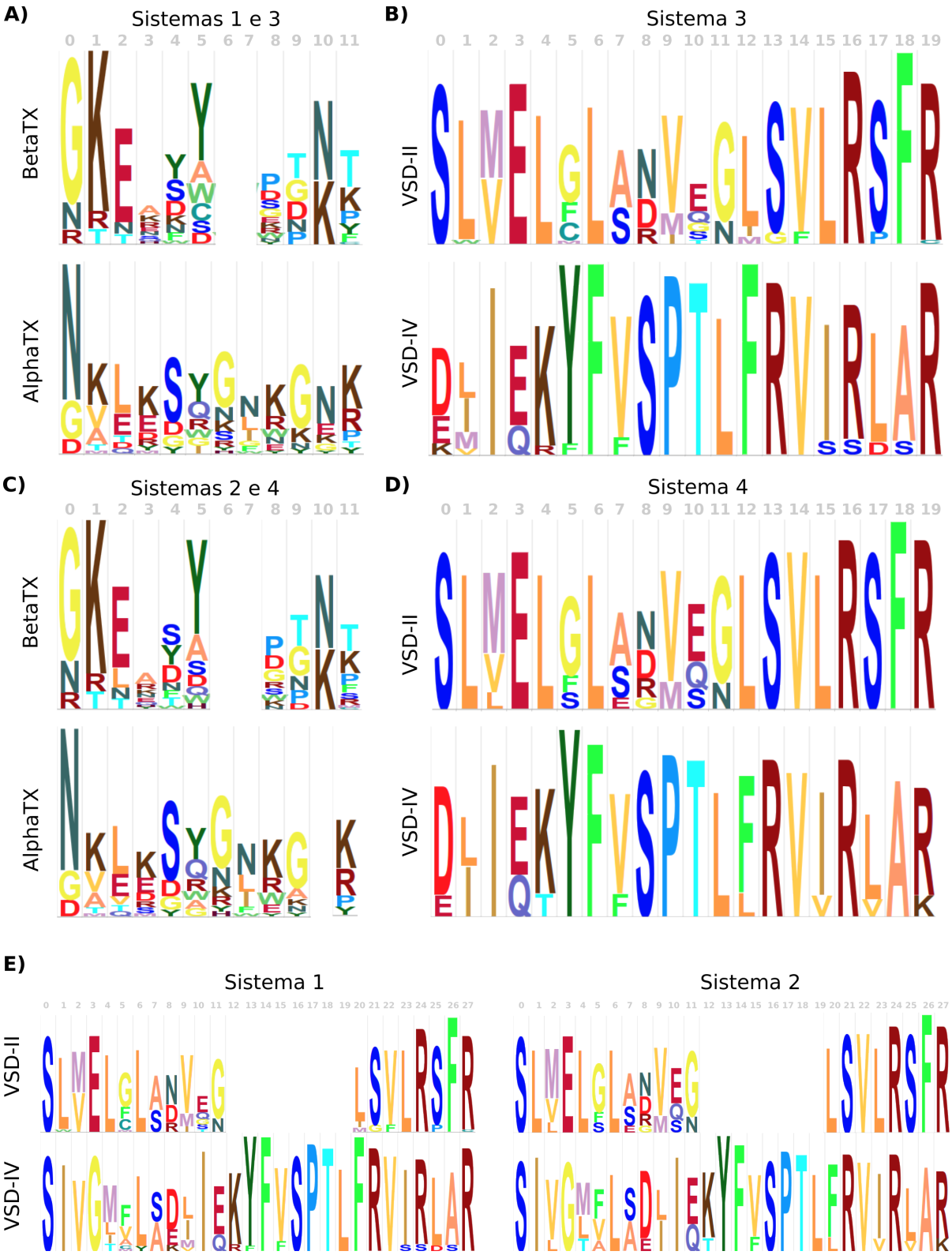

Figura 16: Perfil de aminoácidos dos canais de informação das toxinas em (A) sistemas 1 e 3 e (C) sistemas 2 e 4, e dos VSDs em (B) sistema 3, (D) sistema 4 e (E) sistemas 1 e 2.

Algumas das interações toxina-VSD que fazem parte dos MOC – sistema 1 e MOC – sistema 3 são mostradas nas Tabela 12 e Tabela 13. A Tabela 13 mostra as interações entre toxinas de escorpião e canais Na_V de humanos, enquanto a Tabela 12 mostra as melhores interações toxina-VSD que aparecem nos modelos.

A Tabela 14 mostra as melhores interações toxina-VSD extraídas dos MOCs obtidos pelo AG para os sistemas 2 e 4. A melhor interação para uma β-toxina se deu com $hNa_V1.5$ no sistema 2 e com $hNa_V1.3$ no sistema 4. A melhor interação para uma α-toxina foi com $hNa_V1.1/1.2/1.3/1.6$ no sistema 2 e $hNa_V1.7$ no sistema 4.

A Tabela 15 mostra as energias médias de interação de toxinas com cada tipo de canal Na_V ($hNa_V1.1$-1.7 e Na_V de *Drosophila melanogaster*) extraídas dos MOC – sistema 2 e MOC – sistema 4. Para as β-toxinas, as melhores interações se deram, em média, com canais do tipo $hNa_V1.5$. Já para as α-toxinas, as melhores interações se deram, em média, com canais dos tipos $hNa_V1.7$, no sistema 2, e com canais do tipo $hNa_V1.4/1.5$, no sistema 4.

Tabela 12: Melhores interações com VSD-II e VSD-IV retiradas dos MOC – sistema 1 e MOC – sistema 3 para cada tipo de Na_V

Sistema 1		**Sistema 3**		**VSD**	**Canal**
Toxina	**Energia**	**Toxina**	**Energia**		
TdNa9	-19,2	Tt1g	-19,5	II	$Na_V1.1$
Tst1	-16,6	Tz2	-20,2	II	$Na_V1.2$
β-toxin Im-2	-20,6	LmNaTx64.1	-23,1	II	$Na_V1.3$
TdNa7	-19,0	BmKBT-like peptide	-21,4	II	$Na_V1.4$
LmNaTx64.1	-21,7	To11	-21,7	II	$Na_V1.5$
AaHIT4	-19,4	β-toxin Im-2	-20,8	II	$Na_V1.6$
BmKBT-like peptide	-21,2	To15	-20,9	II	$Na_V1.7$
Amm3	-33,7	Os1	-18,7	IV	$Na_V1.1$
BmK AGP-SYPU	-31,6	AaH4	-17,4	IV	$Na_V1.2$
AaH1	-32,6	BmαTx47	-19,4	IV	$Na_V1.3$
BmKBT	-40,2	Aam3	-19,3	IV	$Na_V1.4$
To10	-32,6	AaH2	-19,5	IV	$Na_V1.5$
MeuNaTx-5	-30,5	Lqq5	-19,5	IV	$Na_V1.6$
Lqh6	-40,9	Bot11	-18,7	IV	$Na_V1.7$

Tabela 13: Pares toxina-VSD que envolvem canais hNa_V retirados dos MOC – sistema 1 e MOC – sistema 3

Sistema 1		Sistema 3		VSD	Canal
Toxina	**Energia**	**Toxina**	**Energia**		
LmNaTx34.5	-16,3	BmKIT4	-16,1	II	$hNa_V1.1/1.2$
Td6	-18,0	BmKAs1	-17,5	II	$hNa_V1.3$
CngtIII	-15,4	TdNa1	-14,6	II	$hNa_V1.4$
LqhIT1d	-17,7	Bactridin-2	-21,7	II	$hNa_V1.5$
BmKAS	-18,5	LmNaTx34.2	-16,3	II	$hNa_V1.6$
Ts1	-14,1	Cex11	-15,5	II	$hNa_V1.7$
BmK-M7	-31,3	MeuNaTx-5	-18,7	IV	$hNa_V1.1/1.2/1.3/1.6$
To10	-32,6	To10	-19,0	IV	$hNa_V1.4/1.5$
Cll9	-27,0	Cll9	-18,4	IV	$hNa_V1.7$

Tabela 14: Melhores interações com VSD-II e VSD-IV retiradas dos MOC – sistema 2 e MOC – sistema 4 para cada tipo de canal Na_V

Sistema 2		Sistema 4		VSD	Canal
Toxina	**Energia**	**Toxina**	**Energia**		
CsEv1	-15,4	BmKAs1	-19,5	II	$hNa_V1.1/1.2$
Css6	-16,8	LmNaTx64.1	-24,8	II	$hNa_V1.3$
BmKAS	-20,9	BmKBT-like peptide	-22,6	II	$hNa_V1.4$
LmNaTx64.1	-24,7	To11	-22,7	II	$hNa_V1.5$
Tpa2	-18,7	Css2	-17,8	II	$hNa_V1.6$
BmKBT-like peptide	-22,3	Cll8	-16,5	II	$hNa_V1.7$
Lqh6	-44,5	BmαTx47	-20,1	IV	$hNa_V1.1/1.2/1.3/1.6$
Makatoxin-3	-37,1	AaH2	-20,3	IV	$hNa_V1.4/1.5$
Tb3	-33,9	CsE9	-21,1	IV	$hNa_V1.7$

As sequências de VSD-IV de $hNa_V1.1$, $hNa_V1.2$, $hNa_V1.3$ e $hNa_V1.6$ são idênticas, no contexto de um alfabeto reduzido de aminoácidos e, portanto, são mostradas nas tabelas de forma agrupada. O mesmo ocorre com as sequências de VSDIV de $hNa_V1.4$ e $hNa_V1.5$, e VSDII de $hNa_V1.1$ e $hNa_V1.2$.

Tabela 15: Energias médias de interação com cada tipo de canal Na_V *retiradas dos MOC – sistema 2 e MOC – sistema 4; N representa o número de canais* Na_V *de cada tipo que existe no MSA*

VSD	Canal	N	Energia média	
			Sistema 2	Sistema 4
II	$hNa_V1.1/1.2$	20	-15,4 ± 0,3	-17,3 ± 2,4
II	$hNa_V1.3$	20	-15,5 ± 1,7	-21,2 ± 3,1
II	$hNa_V1.4$	20	-16,8 ± 4,7	-17,9 ± 4,7
II	$hNa_V1.5$	20	-20,9 ± 6,8	-21,8 ± 0,6
II	$hNa_V1.6$	20	-17,5 ± 0,4	-17,0 ± 0,6
II	$hNa_V1.7$	20	-16,5 ± 3,2	-16,0 ± 0,4
II	Na_V *D. melanogaster*	11	-17,2 ± 1,9	-17,5 ± 0,3
IV	$hNa_V1.1/1.2/1.3/1.6$	38	-34,6 ± 13,3	-17,8 ± 1,9
IV	$hNa_V1.4/1.5$	40	-32,9 ± 3,4	-18,7 ± 0,6
IV	$hNa_V1.7$	20	-34,6 ± 5,1	-17,9 ± 3,3
IV	Na_V *D. melanogaster*	09	-33,9 ± 0,0	-18,3 ± 1,4

Análise Estatística das Interações entre Canais que Caracterizam o Modelo Otimizado de Coevolução

Os resultados que serão mostrados nesta sessão são referentes à análise das energias de interação entre canais de informação das toxinas e canais de informação dos VSDs. O vetor médio das energias de interação guarda informação sobre qual é a energia média de interação entre os canais i (da toxina) e j (do VSD) para todos os pares de canais (i, j) possíveis. Esse vetor é obtido calculando-se a média, posição por posição, dos vetores que guardam as energias de interação, pesadas pelo acoplamento entre os resíduos, de cada par de aminoácidos que pode estar em contato quando uma determinada toxina interage com o VSD pelo qual ela possui maior afinidade, segundo o MOC.

A Figura 17 mostra a projeção do vetor médio das energias de interações nas 50 primeiras da PCA para o conjunto de pares formados com VSD-II (em vermelho) e para o conjunto de pares formados com VSD-IV (em preto) nos MOC – sistema 1-4. Os dois primeiros autovetores explicam, respectivamente, 52% e 20% da variância dos dados para os sistemas 1 e 2, 61% e 16% da variância dos dados para o sistema 3, e 60% e 18% da variância dos dados para o sistema 4. Pode-se notar que ambos os conjuntos tiveram projeção positiva

na primeira componente encontrada, porém, na segunda componente, a projeção do conjunto VSD-IV foi positiva, enquanto a projeção do conjunto VSD-II foi negativa, ou vice-versa.

As Figuras 18–21 mostram as projeções das energias de cada par toxina-VSD do conjunto que interage com VSD-IV e do conjunto que interage com VSD-II no espaço formado pelas duas primeiras componentes da PCA, sendo a primeira componente (v_0) representada no eixo y e a segunda componente (v_1) no eixo x. As projeções são mostradas antes (em vermelho e preto) e depois (em azul e verde) da remoção de um determinado conjunto de interações. As interações removidas na letra (B) deslocam a nuvem preta no sentido negativo de v1; já as interações removidas na letra (C), deslocam a nuvem vermelha no sentido positivo de v1; e, na letra (D), as interações removidas deslocam as duas nuvens no sentido negativo de v0. Em todos os casos, os deslocamentos observados são estatisticamente significativos ($p < 0{,}0001$). A letra (E) mostra quais interações entre canais foram removidas para se obter os efeitos observados em (B), (C) e (D) nas cores preta, vermelha e azul, respectivamente. As Figuras 18–21 são referentes aos MOC – sistema 1–4, respectivamente.

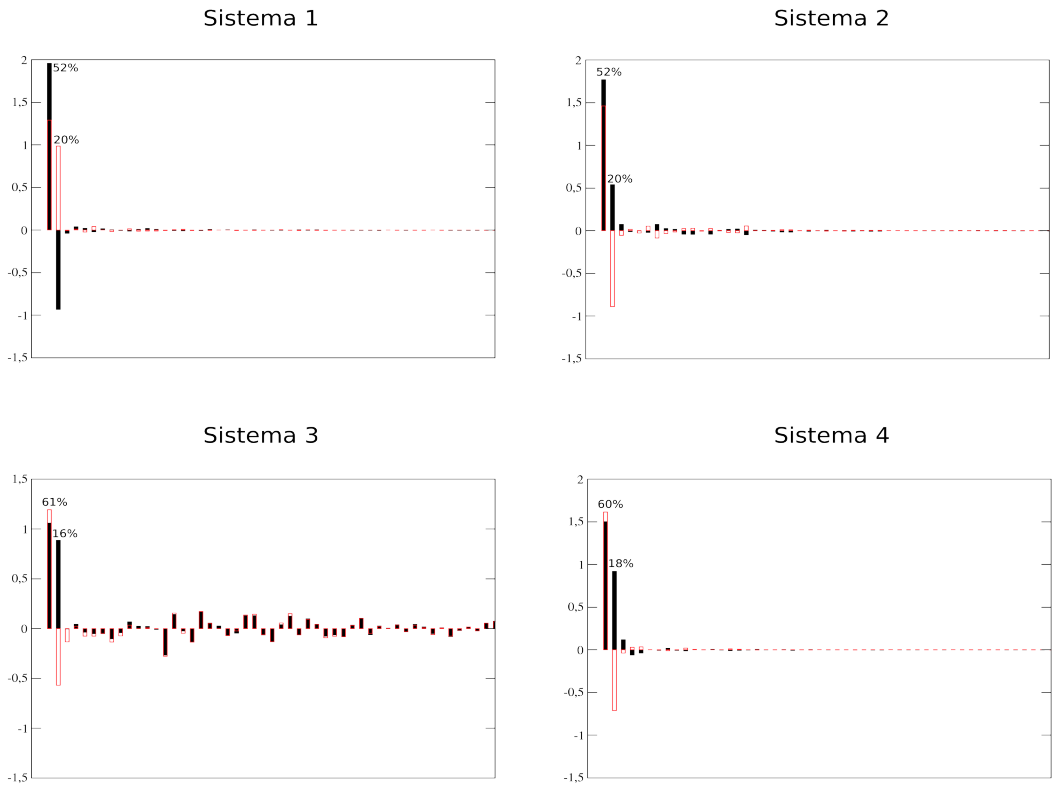

Figura 17: Projeção das energias médias de α-toxinas (em preto) e β-toxinas (em vermelho) nas 50 primeiras componentes que explicam a variância das energias obtidas do MOC para os sistemas 1, 2, 3 e 4. Acima das projeções nas componentes 1 e 2, está indicada a contribuição daquela componente para explicar a variância das energias do MOC.

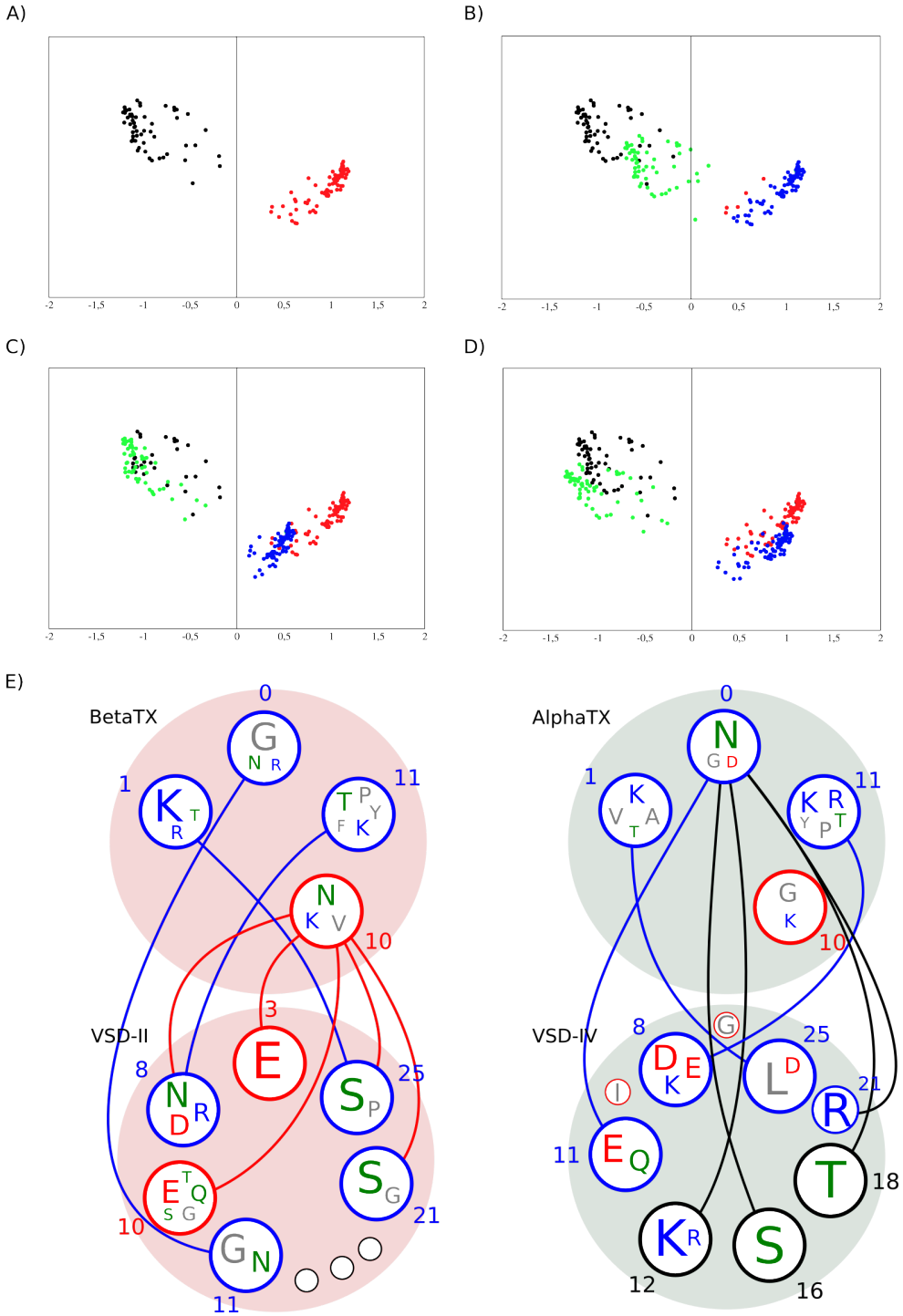

Figura 18: Interações que determinam a afinidade diferencial de α- e β-toxinas por VSD-IV e -II encontradas a partir da PCA feita no MOC – sistema 1. A) Projeções das energias de cada par tox-VSDII (n = 121), em vermelho, e tox-VSDIV (n = 84), em preto, no espaço gerado pelos vetores V0 e V1, que explicam, respectivamente, 52% e 20% da variância dos dados. B, C e D: Projeções das energias dos pares antes, em vermelho e preto, e depois, em azul e verde, da remoção das interações com projeção > 0,05 no vetor V1 (B), da remoção das interações com projeção < -0,05 no vetor V1 e da remoção das interações com projeção > 0,05 no vetor V0. E) Representação das interações que foram excluídas em B, em preto, em C, em vermelho e em D, em azul.

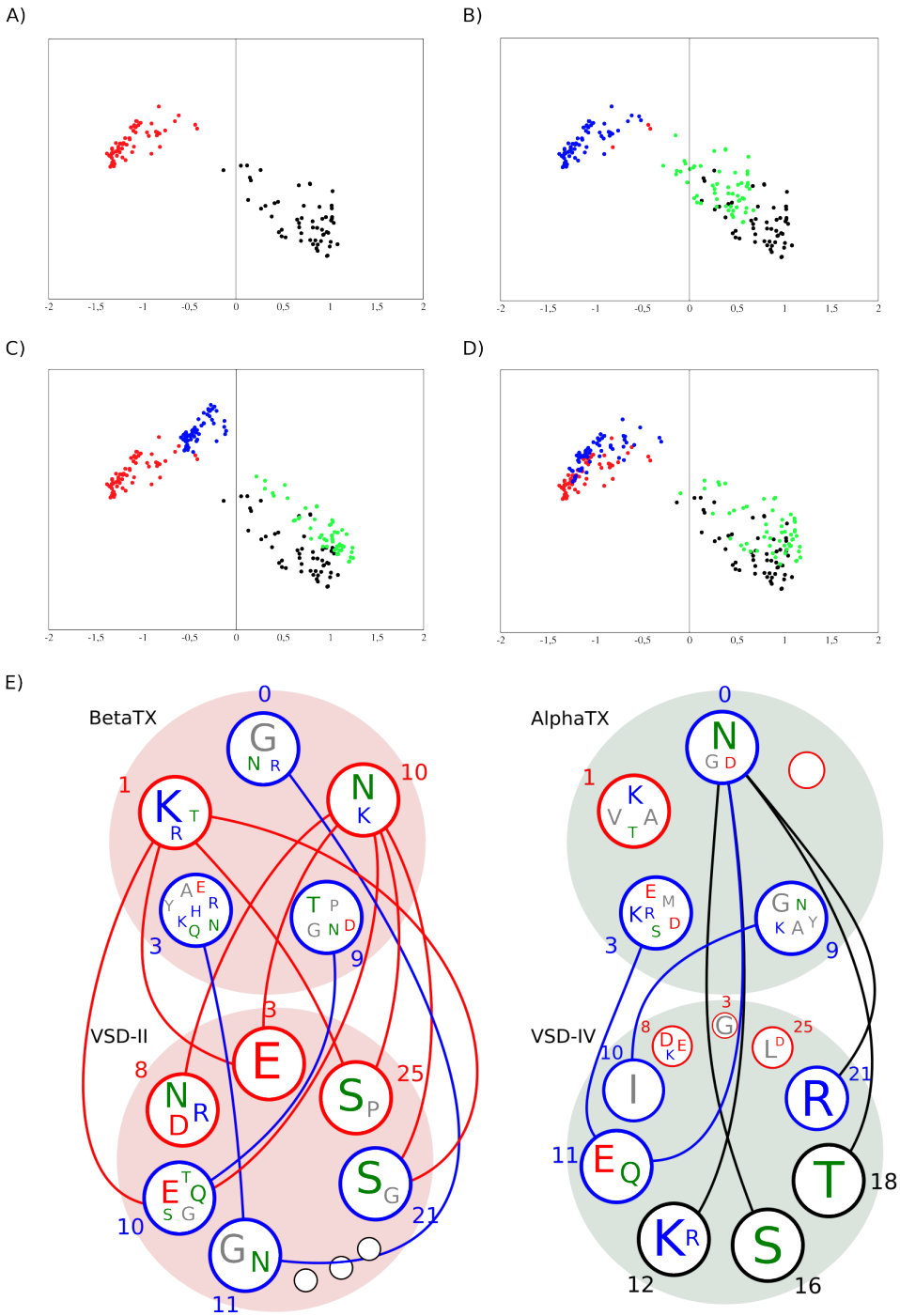

Figura 19: Interações que determinam a afinidade diferencial de α- e β-toxinas por VSD-IV e -II encontradas a partir da PCA feita no MOC – sistema 2. A) Projeções das energias de cada par tox-VSDII (n = 121), em vermelho, e tox-VSDIV (n = 84), em preto, no espaço gerado pelos vetores V0 e V1, que explicam, respectivamente, 52% e 20% da variância dos dados. B, C e D: Projeções das energias dos pares antes, em vermelho e preto, e depois, em azul e verde, da remoção das interações com projeção > 0,05 no vetor V1 (B), da remoção das interações com projeção < -0,05 no vetor V1 e da remoção das interações com projeção > 0,05 no vetor V0. E) Representação das interações que foram excluídas em B, em preto, em C, em vermelho e em D, em azul.

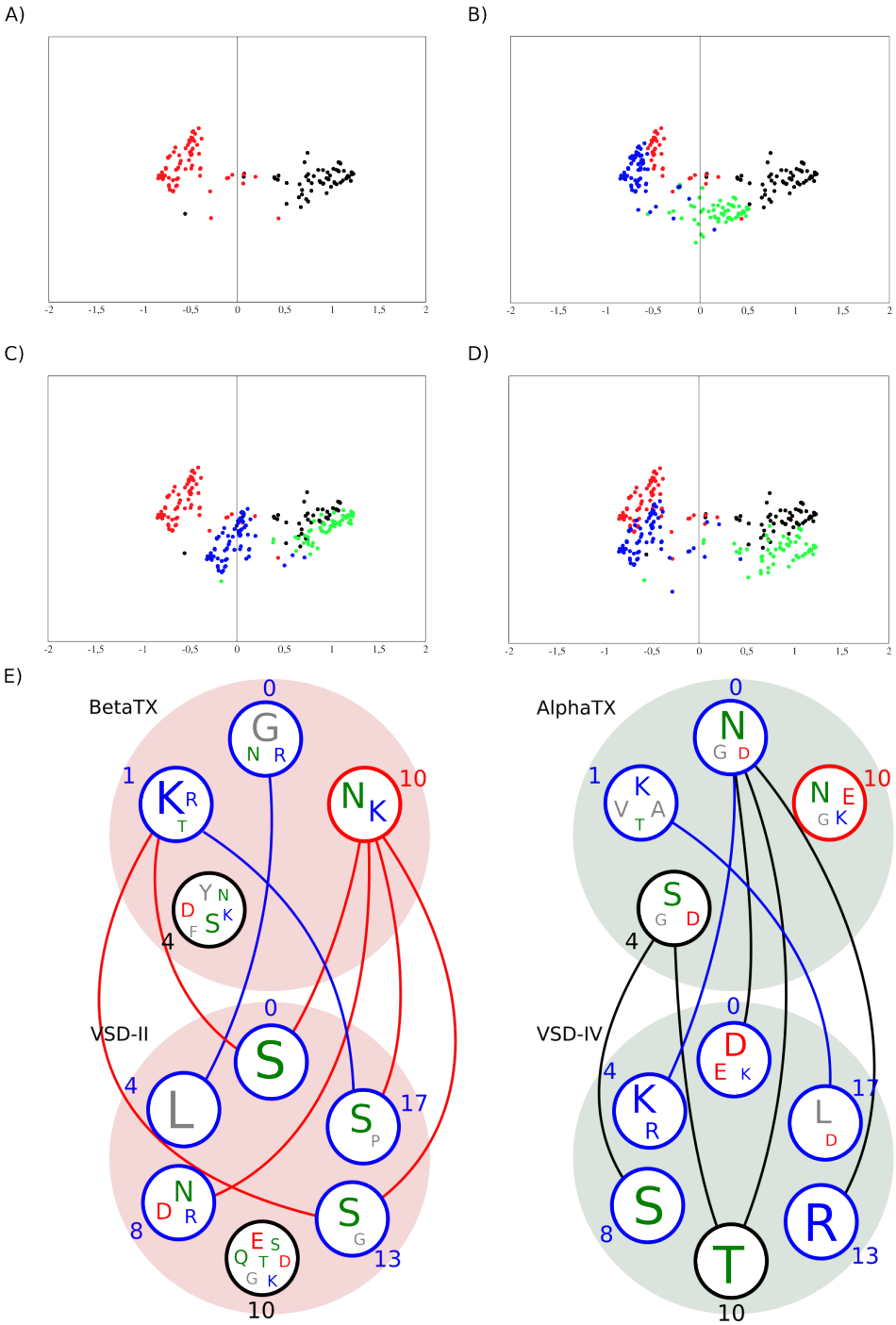

Figura 20: Interações que determinam a afinidade diferencial de α- e β-toxinas por VSD-IV e -II encontradas a partir da PCA feita no MOC – sistema 3. A) Projeções das energias de cada par tox-VSDII (n = 121), em vermelho, e tox-VSDIV (n = 84), em preto, no espaço gerado pelos vetores V0 e V1, que explicam, respectivamente, 61% e 16% da variância dos dados. B, C e D: Projeções das energias dos pares antes, em vermelho e preto, e depois, em azul e verde, da remoção das interações com projeção > 0,05 no vetor V1 (B), da remoção das interações com projeção < -0,05 no vetor V1 e da remoção das interações com projeção > 0,05 no vetor V0. E) Representação das interações que foram excluídas em B, em preto, em C, em vermelho e em D, em azul.

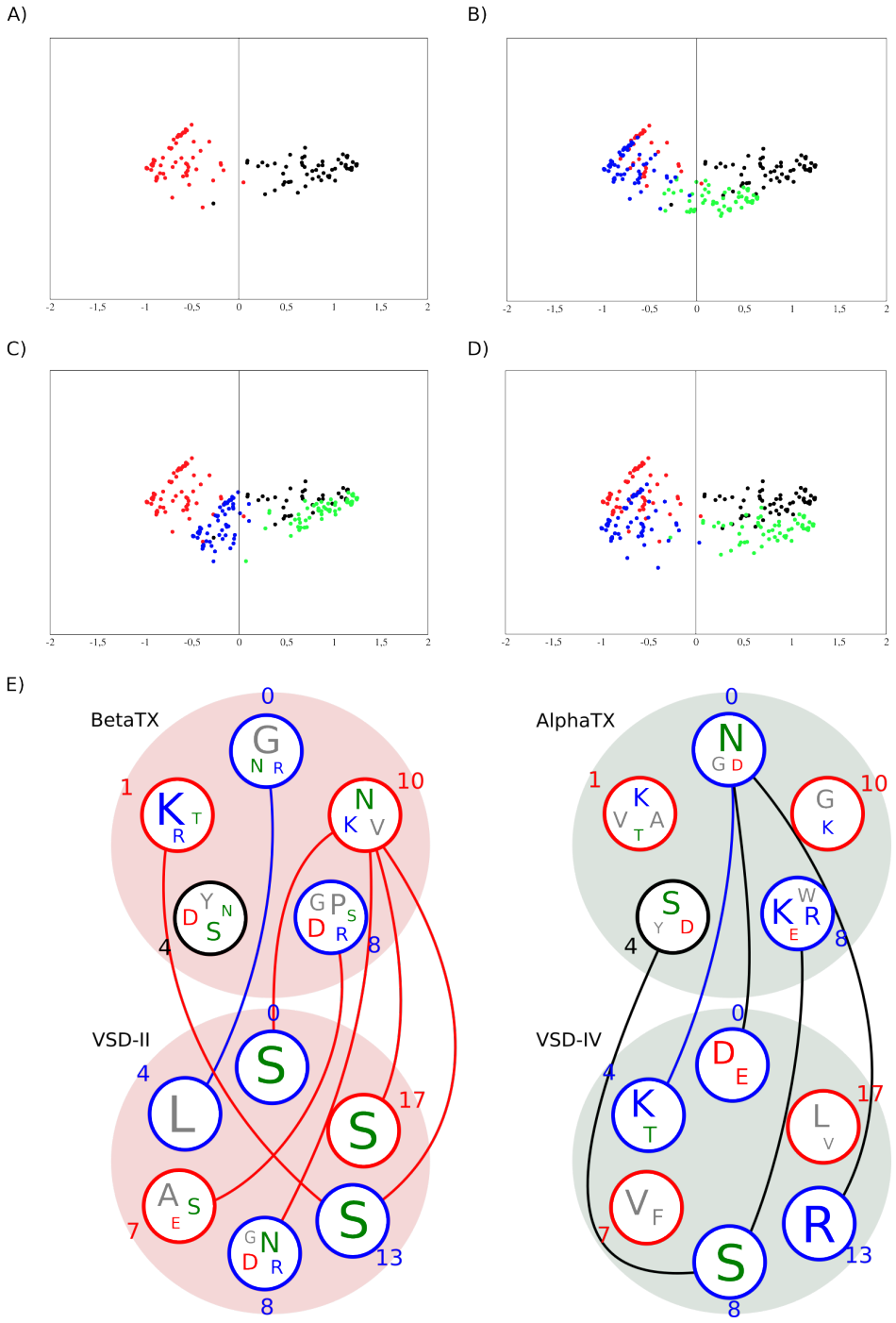

Figura 21: Interações que determinam a afinidade diferencial de α- e β-toxinas por VSD-IV e -II encontradas a partir da PCA feita no MOC – sistema 4. A) Projeções das energias de cada par tox-VSDII (n = 131), em vermelho, e tox-VSDIV (n = 107), em preto, no espaço gerado pelos vetores V0 e V1, que explicam, respectivamente, 60% e 18% da variância dos dados. B, C e D: Projeções das energias dos pares antes, em vermelho e preto, e depois, em azul e verde, da remoção das interações com projeção > 0,05 no vetor V1 (B), da remoção das interações com projeção < -0,05 no vetor V1 e da remoção das interações com projeção > 0,05 no vetor V0. E) Representação das interações que foram excluídas em B, em preto, em C, em vermelho e em D, em azul.

A matriz de distâncias de Hamming par a par para o MSA de α- e β-toxinas é mostrada na Figura 22. As α- e β-toxinas com estrutura 3D, e algumas β-toxinas com função determinada experimentalmente, mas sem estrutura 3D aparecem marcadas na diagonal.

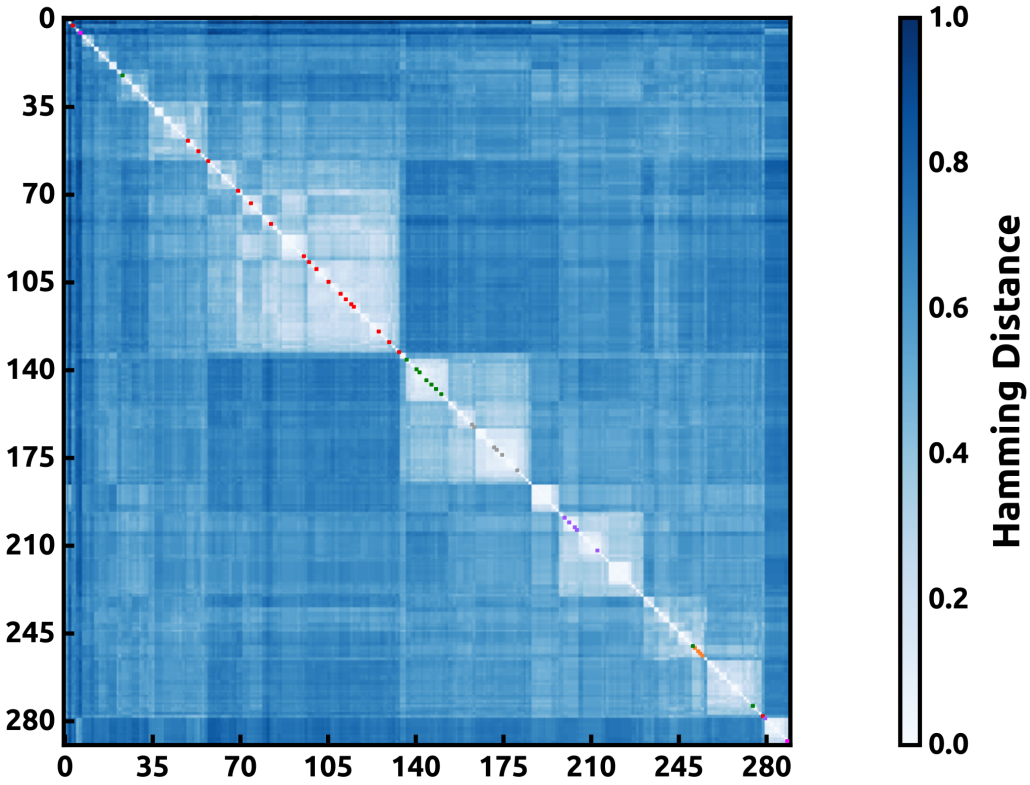

Figura 22: α- e β-toxinas agrupadas segundo critério de distância de Hamming. As α- e β-toxinas com estrutura 3D, e algumas β-toxinas com função determinada experimentalmente, mas sem estrutura 3D aparecem em destaque na diagonal. As α-toxinas estão marcadas em vermelho e as β-toxinas sem anotação estão marcadas em cinza, β-toxinas anti-mamífero em verde, anti-inseto/mamífero em laranja e anti-inseto em roxo/rosa.

Foram simuladas 128 α- e β-toxinas, porém, devido à necessidade de refinamento dos cálculos, algumas das toxinas foram retiradas do banco de dados inicial e serão analisadas posteriormente. Após a seleção das toxinas que continuariam a compor o banco, o número total de elementos foi reduzido para 82 (Anexo 4). O RMSD médio do banco é mostrado na Figura 23 e o perfil de acessibilidade médio na Figura 24.

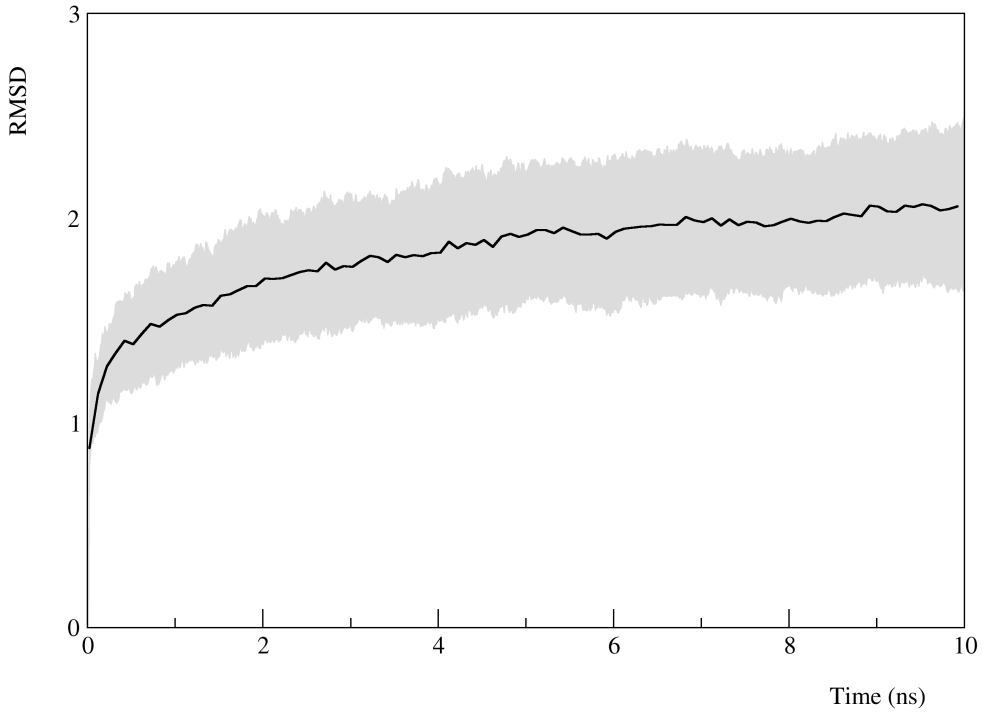

Figura 23: Média da variação de RMSD dos átomos pesados de α- e β- toxinas ao longo do tempo de simulação (linha preta; n = 82). A área em cinza representa a variância.

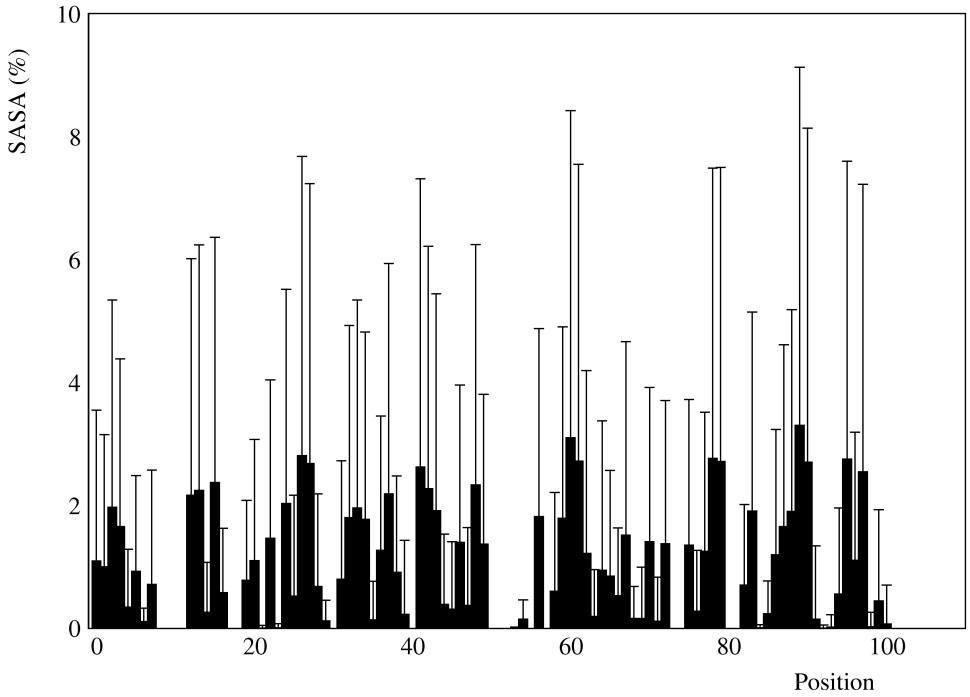

Figura 24: Perfil de acessibilidade médio de α- e β-toxinas do banco (n = 82). Foi calculada a média dos SASA por posição do MSA, com base no perfil de SASA obtido para cada toxina ao longo da simulação. As barras de erro representam o desvio padrão.

Discussão

Canais de Informação no MSA de Toxinas

Pode-se observar na Tabela 9 que 10 dos 12 canais de informação encontrados são conservados nas α-toxinas, enquanto apenas seis são conservados nas β-toxinas. Esse resultado pode ser devido ao fato de que as α-toxinas são mais parecidas entre si do que as β-toxinas, como mostra a Figura 22, na qual as α-toxinas formam um *cluster* mais bem-definido do que as β-toxinas, que formam *clusters* separados, de acordo com o subtipo.

Quando as posições conservadas (Tabela 8) e os canais de informação (Tabela 9) são mapeados em algumas α- e β-toxinas (Figura 25 e Figura 26, respectivamente), pode-se ver o nítido contraste entre o núcleo hidrofóbico conservado e a superfície funcional hidrofílica. Nas estruturas à esquerda, são mostrados os aminoácidos conservados em α- e β-toxinas (cor sólida) e a superfície total da toxina (translúcida). Nas estruturas à direita, são mostrados os aminoácidos que tendem a ser conservados em um grupo, mas diferentes entre os grupos.

O fato de os canais de informação, determinados através da análise informacional do alinhamento de sequências primárias, se encontrarem na superfície das toxinas é coerente com a hipótese de que esses canais de informação fazem parte da superfície de interação de α- e β-toxinas com os canais Na_V. Há muitas evidências experimentais que corroboram essa hipótese.

Os trabalhos das equipes de Kahn, Liu, Gur e Wang corroboram a hipótese de que as posições 67, 68 e 88, que correspondem aos resíduos G43, N44 e K58 na α-toxina Lqh2, fazem parte da superfície funcional de α-toxinas (Gur et al., 2011; Kahn et al., 2009; Liu et al., 2005; Wang et al., 2011). Kahn e colaboradores também encontram em seu trabalho evidências de que o aminoácido K62 de Lqh2 (posição 96) seria importante para a ligação dessa toxina com o canal (Kahn et al., 2009).

Os trabalhos das equipes de Cohen e Cestèle corroboram a hipótese de que as posições 39, 43, 60 e 88, correspondentes aos resíduos E28, Q32, Y42 e W58 em Css4, fazem parte da superfície funcional de β-toxinas (Cestèle et al., 2006; Cohen et al., 2004; Pedraza Escalona and Possani, 2013). Já as posições 17 e 21 parecem ser especialmente importantes para a ligação com canais iônicos de insetos (Hassani et al., 1999; Pedraza Escalona and Possani, 2013).

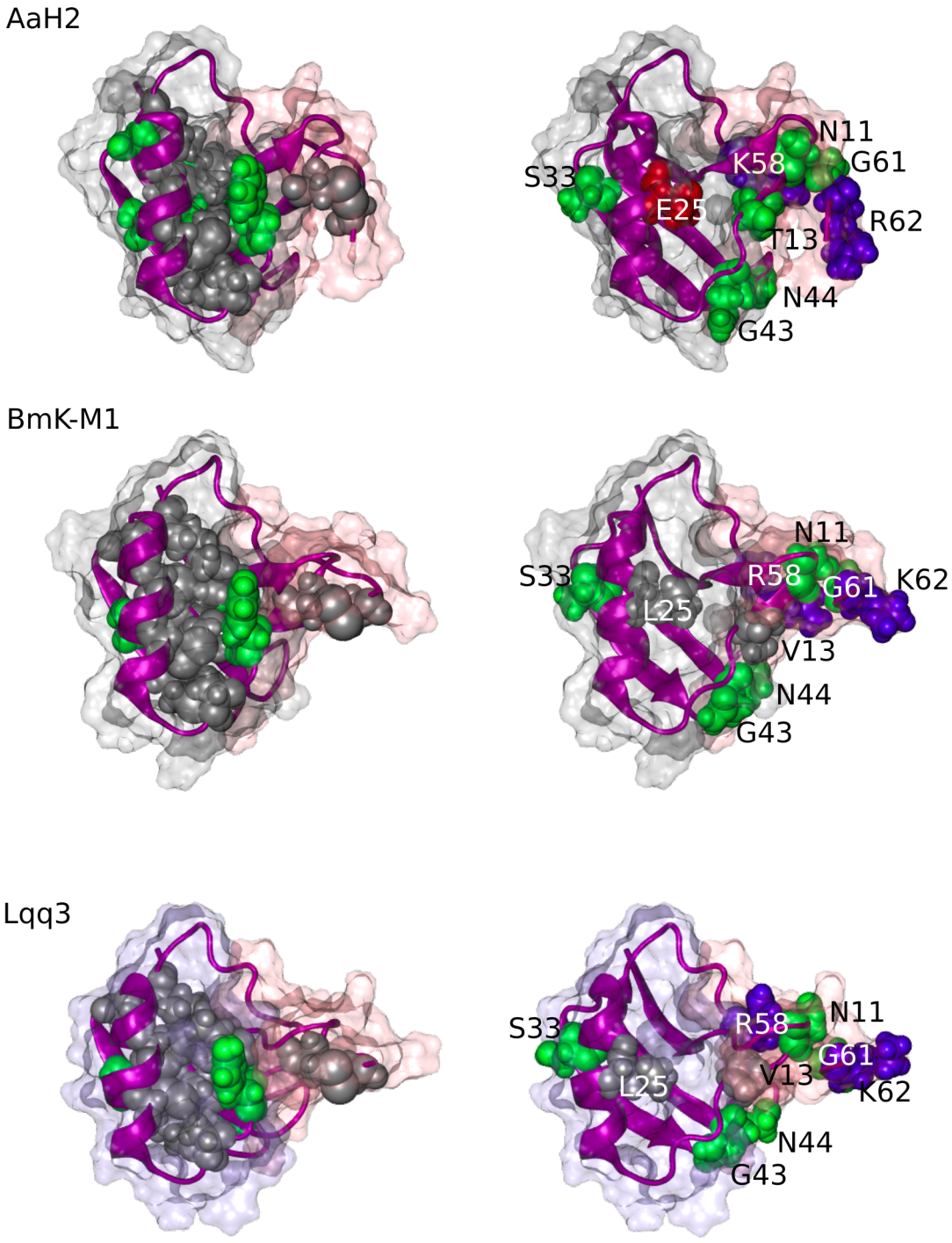

Figura 25: α-toxinas de escorpião de diferentes classes: BmK-M1, toxina anti-inseto; Lqq3, toxina α-like; e Aah2, toxina anti-mamífero. As estrutura da esquerda apresentam em destaque os aminoácidos de posições consevadas em α- e β-toxinas (H < 0,5), segundo análise de entropia. As estruturas da direita apresentam em destaque aminoácidos de posições que são conservadas em α-toxinas, mas diferentes em α- e β-toxinas, segundo análise de informação mútua. Resíduos apolares estão coloridos em cinza, resíduos polares em verde, resíduos de carga positiva em azul e resíduos de carga negativa em vermelho. O domínio NC está representado em rosa e o domínio do núcleo em cinza.

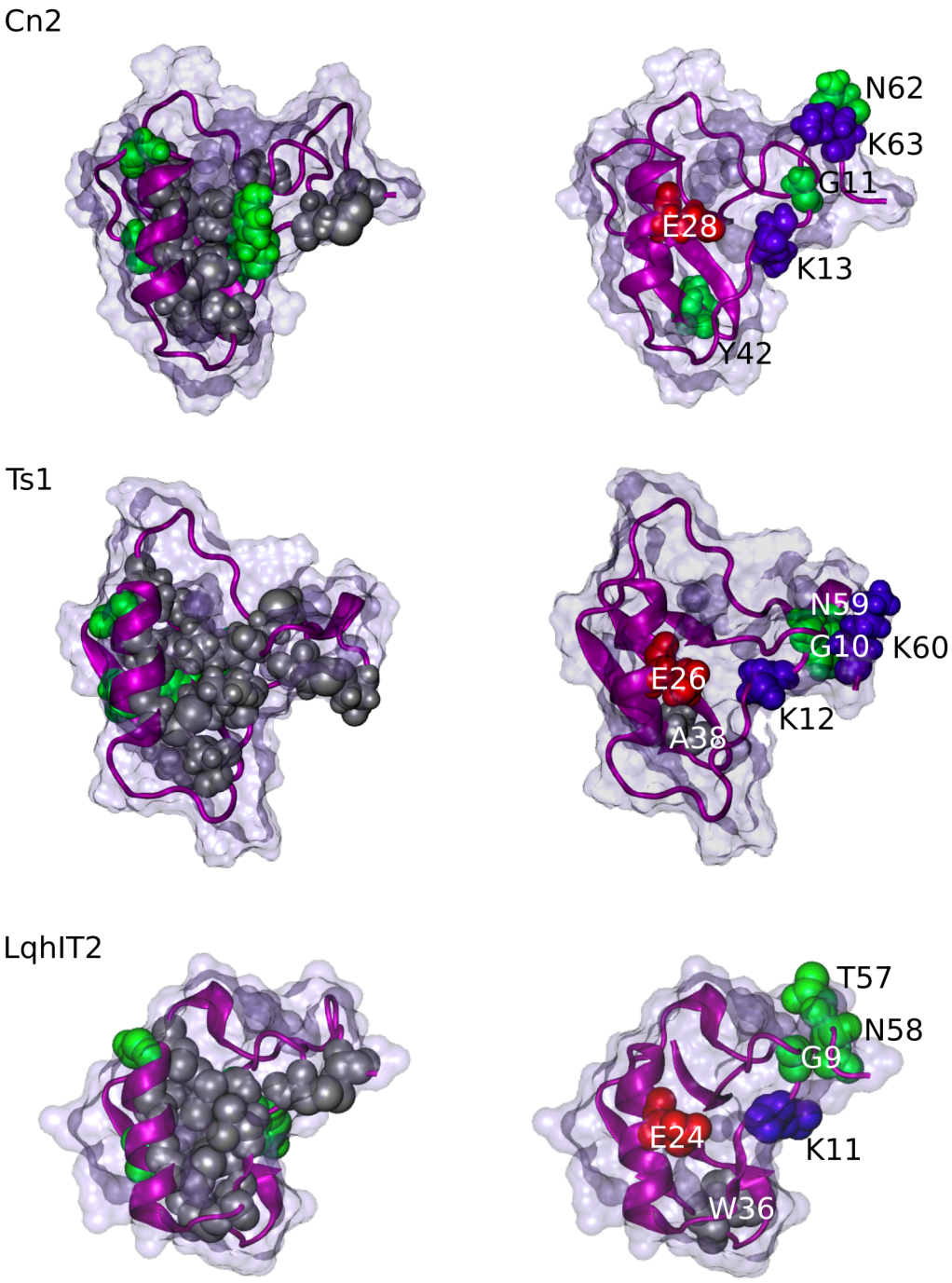

Figura 26: β-toxinas de escorpião de diferentes classes: Cn2, toxinas anti-mamífero; LqhIT1, toxina anti-inseto; e Ts1, toxina anti-mamífero/inseto. As estrutura da esquerda apresentam em destaque os aminoácidos de posições conservadas em α- e β-toxinas (H < 0,5), segundo análise de entropia. As estruturas da direita apresentam em destaque aminoácidos de posições que são conservadas em β-toxinas, mas diferentes em α- e β-toxinas, segundo análise de informação mútua. Resíduos apolares estão coloridos em cinza, resíduos polares em verde, resíduos de carga positiva em azul e resíduos de carga negativa em vermelho.

Experimentos de mutagênese apontam também alguns dos aminoácidos conservados em α- e β-toxinas como importantes para sua função, entretanto, tais posições não seriam interessantes no contexto deste trabalho, pois não guardam informação de acoplamento com os canais de informação dos VSD de canais Na_V.

Os resultados das análises com diferentes conjuntos de canais de informação (Figura 13 e Figura 14) deixam claro que o conjunto encontrado utilizando TI é uma boa escolha, pois, quanto mais distante desse conjunto, maiores são as energias dos MOCs encontrados após 5.000 gerações do AG. Além disso, na Figura 14 fica evidente a diminuição do *gap* entre as energias aleatórias e a energia do MOC, à medida em que o conjunto de canais de informação sofre perturbações cada vez maiores.

Modelo Otimizado de Coevolução

Na sessão de Resultados, foram mostrados os resultados obtidos para os sistemas 1, 2, 3 e 4. Os sistemas 1 e 2 contém um alinhamento de VSDs gerado com o programa ClustalX, enquanto os sistemas 3 e 4 contém um alinhamento de VSDs baseado em um robusto alinhamento de VSDs feito por Palovcak e colaboradores (Palovcak et al., 2014). O fato de existirem dois alinhamentos possíveis para essa região levanta questões sobre como os *loops* extracelulares de canais Na_V estariam estruturados e se existiriam diferenças em termos da quantidade de aminoácidos nos *loops* S3-S4 de VSDs lentos (IV) e rápidos (I, II e III). Pelo alinhamento utilizado nos sistemas 1 e 2, seria de se supor que o *loop* S3-S4 do domínio IV é maior que o *loop* S3-S4 do domínio II. Segundo o alinhamento dos sistemas 3 e 4, entretanto, a diferença de tamanho estaria entre as hélices S3 dos domínios II e IV.

Comparando-se as Tabelas 12–15 fica evidente que cada sistema apresentou resultados bastante distintos em termos de pares de maior afinidade. Entretanto, pode-se notar que alguns dos pares toxina-VSD de maior afinidade são os mesmos para os sistemas 3 e 4, a saber: LmNaTX64.1 e Nav1.3, *BmKBT-like peptide* e Nav1.4, e To11 e Nav1.5. Além disso, os pares de maior afinidade To10–NaV1.4/1.5 e Cll9–NaV1.7 se mantém para os sistemas 2 e 4. Esses resultados podem estar indicando uma relativa similaridade entre as interações com VSD-II nos MOCs dos sistemas 3 e 4 e entre as interações com VSD-IV nos MOCs dos sistemas 2 e 4.

Existem cinco toxinas que possuem testes de afinidade contra todos os canais hNav publicados na literatura nos trabalhos de Schiavon et al. e Peigneur et al. (Peigneur et al.,

2015; Schiavon et al., 2012). Quando os dados experimentais são comparados aos dados obtidos do MOC – sistema 4 (Figura 27), a melhor taxa de acerto, de 3,5 em 5 (contando um par de afinidade experimental moderada como 0,5 acerto) é obtida, o que indica que o modelo é capaz de recuperar pelo menos uma parcela das afinidades reais. O fato de que nem todas as afinidades são recuperadas de forma correta pode ser por conta do número limitado de VSDs de Na_Vs de cada tipo (1.1 a 1.7) disponível para interação com as toxinas. A toxina Cll1, como não é muito específica para nenhum tipo de Na_V, pode não ter tido a capacidade de competir com as outras toxinas por algum outro tipo de Na_V e acabou tendo que se ligar ao canal $Na_V1.7$, o qual, aparentemente, não é o alvo preferido de quase nenhuma toxina β.

É interessante notar que várias das β-toxinas anotadas como toxinas anti-mamífero, que possuem uma glutamina (Q) no CI 3 (Q32 em Cn2), formaram pares de melhor afinidade com canais hNav1.6 no MOC – sistema 4, indicando uma possível correlação entre esse CI e o CI 5 dos VSDs (S833 em hNav1.6), o qual aparece carregado exclusivamente nesse tipo de canal. Os canais hNav1.1-1.5 e hNav1.7 apresentam uma glicina nesse CI, a qual pode estar interagindo com os resíduos hidrofóbicos que aparecem no CI 3 das outras toxinas.

De fato, há trabalhos que mostram que as toxinas pertencentes à classe de toxinas beta anti-mamíferos interagem principalmente com canais dos tipos $Na_V1.1$, $Na_V1.2$, $Na_V1.4$ and $Na_V1.6$ e não apresentam efeito em canais do tipo $Na_V1.5$ (Cestèle et al., 1998; Marcotte et al., 1997; Pedraza Escalona and Possani, 2013).

As toxinas que formaram pares de melhor afinidade com canais hNav1.5 são, em sua maioria, produzidas por escorpiões da espécie *Tityus discrepans* ou da espécie *Rhopalurus junceus*.

Observando a Tabela 15, pode-se notar quatro padrões distintos em relação à energia média de interação com um determinado tipo de canal hNa_V: (1) energia média abaixo da média geral e variância < 1,0 ($hNa_V1.5$–DII e $hNa_V1.4/1.5$–DIV), (2) energia média abaixo da média geral e variância > 1,0 ($hNa_V1.3$–DII), (3) energia média acima da média geral e variância < 1,0 ($hNa_V1.6/1.7$–DII), e (4) energia média acima da média geral e variância > 1,0 ($hNa_V1.1/1.2/1.4$–DII e $hNa_V1.1/1.2/1.3/1.6/1.7$–DIV).

	hNa$_V$1.1	**hNa$_V$1.2**	**hNa$_V$1.3**	**hNa$_V$1.4**	**hNa$_V$1.5**	**hNa$_V$1.6**	**hNa$_V$1.7**
Ts1	-	-	+	+	++	++	-
Cn8	++	++	+	++	++	++	-
Css2	-	-	-	-	-	++	-
Cll2	++	-	-	+	+	++	-
Cll1	++	++	++	++	++	+	-

Figura 27: Matriz de afinidades teóricas e experimentais. Os resultados obtidos através de experimentos de eletrofisiologia estão marcados com '+' para afinidade boa, '++' para afinidade ótima, e '-' para não-afinidade. Os resultados teóricos obtidos do MOC – sistema 4 para pares de toxinas de mamífero e canais hNa$_V$1.1-1.7 estão coloridos em azul, quando correspondem a uma afinidade experimental ótima, em amarelo, quando correspondem a uma afinidade experimental boa e em vermelho quando correspondem a uma não-afinidade experimental.

É difícil determinar o significado de cada um dos padrões acima, entretanto, é interessante notar que, para os canais hNa$_V$1.7–DII, foi encontrada a pior energia média de interação e variância baixa, o que pode significar que esse tipo de canal está super-representado no sistema, ou seja, algumas das toxinas que se ligam a ele, poderiam apresentar energias de interação melhores com outros tipos de canal. Essa hipótese é corroborada pelo fato de que, para as toxinas beta testadas em canais humanos por meio de experimentos de eletrofisiologia, não foram encontradas boas afinidades por canais hNa$_V$1.7.

Interações que Definem as Regras do Modelo Otimizado de Coevolução

Foram encontradas, em todos os sistemas, correlações entre o CI 1 das toxinas, correspondente ao resíduo K12 em Ts1, e alguns CIs dos VSDs. A importância do resíduo K12, que forma uma cavidade com o resíduo W54, para a ligação das β-toxina anti-mamífero/inseto foi confirmada no trabalho de Pedraza Escalona e Possani (Pedraza Escalona and Possani, 2013). O resíduo N59 (CI 10 das toxinas), mencionado como parte da superfície funcional de β-toxinas anti-inseto, também forma uma cavidade com o W54, o que leva à hipótese de que possa haver uma relação entre os resíduos K12, W54 e N59 em Ts1 e que a tríade contribua para a interação com o sensor de voltagem. Considerando tal hipótese, os resultados encontrados no presente trabalho em relação à superfície de interação das β-toxinas com o canal também são coerentes.

Nos MOCs dos sistemas 1 e 2 há correlação de CIs 0 e 3 das toxinas com o CI 11 do

VSD-II, que corresponde ao resíduo G845 em $hNa_V1.2$, essencial para a ligação de toxinas β com o sítio 4, conforme mostra a Figura 5. O CI 3 das toxinas corresponde ao resíduo Q32 em Css4, que é importante para a ligação da toxinas, conforme evidências apresentadas em sessões anteriores. O CI 3 dos VSD-II, o qual corresponde ao resíduo E837 em $hNa_V1.2$, comprovadamente necessário para a função da toxina, apresentou correlação com o CI 10 das toxinas, que corresponde uma lisina na região C-terminal de Css4.

Já nos MOCs dos sistemas 3 e 4 foi observada correlação entre o CI 0 das toxinas e o CI 4 dos VSDs, que corresponde ao resíduo L840 em $hNa_V1.2$, comprovadamente necessário para a função da toxinas. As demais correlações encontradas podem ter surgido por conta de interações entre resíduos que ainda não foram identificadas experimentalmente como importantes para a função das toxinas ou podem ser correlações indiretas.

Os MOCs dos diferentes sistemas apresentaram resultados diversos em termos de correlações encontradas por meio da PCA. Entretanto, os resultados foram mais similares entre os sistemas 1 e 2, e entre os sistemas 3 e 4, o que já era esperado, pois esses pares de sistemas foram simulados com o mesmo tipo de alinhamento de VSDs. Não é possível determinar qual dos MOCs obtido representa melhor a realidade, entretanto, visto que as interações do MOC – sistema 4 apresentaram os melhores resultados quando comparadas a afinidades experimentais, é possível que os sistemas 3 e 4 sejam mais realistas.

Em seu trabalho de 2011, Wang et al. afirmam que os aminoácidos T1560, no loop S1-S2 do domínio IV, e E1613 e F1610, no loop S3-S4 do domínio IV, são os constituintes primários do sítio de ligação das α-toxinas, enquanto resíduos do domínio I do canal constituem um sítio de interação secundário (Wang et al., 2011).

Ainda nesse trabalho, Wang et al. desenvolvem um modelo molecular da interação de uma toxina alfa com o com o domínio IV do canal $Na_V1.2$ e discutem que, apesar da posição de *docking* ser parecida com a posição de interação de uma β-toxina com o domínio II (Cestèle et al., 2006), não houve interação de resíduos do canal iônico com resíduos do domínio NC da toxina, que são importantes para a ligação com o VSD.

No presente trabalho, entretanto, foram encontradas evidências de interação de aminoácidos do domínio NC de α-toxinas com o seu sítio primário de ligação no VSD. Esses resultados são esperados por conta do papel fundamental desempenhado por alguns resíduos desse domínio para a função da toxina.

A correlação entre o CI 8 (K58 em AaH II) de α-toxinas e o CI 8 de VSD-IV é parte

das regras que definem o MOC – sistema 4 e parece indicar uma interação importante para a interação de α-toxinas e canais iônicos de humanos. Essa correlação possui respaldo na literatura, nos trabalhos de Kharrat et al. e Legros et al. descritos abaixo.

Kharrat et al. descobriram, em 1989, que o resíduo K58 do C-terminal é indispensável para a função de α-toxinas, o que confirmou a importância da região C-terminal para a interação dessas toxinas com canais Na_V (Kharrat et al., 1989). Por exemplo, em AaH II, a substituição do resíduo K58 por um resíduo hidrofóbico (V e I) ou ácido (E) gera um análogo inativo (Legros et al., 2005). Modificações químicas feitas nas toxinas AaH I, II e III mostram que os resíduos carregados nas regiões N- e C-teminal, bem como os aminoácidos do motivo CSH, do C-terminal e da loop β2–β3 (resíduos 37–44) também são muito importantes para a atividade das α-toxin e para a sua interação com canais Na_V (Kharrat et al., 1989; Martin-Eauclaire et al., 2014).

Também foram encontradas correlações entre o CI 0 (N11 em AaH II) de α-toxinas e diversos resíduos do *loop* extracelular S3-S4 do domínio IV de canais Na_V, o que está de acordo com as conclusões de Chen e Chung, os quais afirmam, em seu trabalho de 2012, que a superfície funcional das α-toxinas anti-mamífero tem como centro o chamado *liker-domain*, constituído pelos resíduos 8-18 (Chen and Chung, 2012).

Granier et al. investigou a interação da toxina AaH II com o seu sítio de ligação no canal Nav utilizando cinco populações de anticorpos selecionados por sua especificidade por várias regiões da toxina AaH II (Granier et al., 1989). Esses estudos indicaram que dois sítios antigênicos estavam envolvidos nos mecanismos moleculares de neutralização de toxicidade. Um deles fica localizado próximo à ponte dissulfeto entre os resíduos Cys12 e Cys63, a qual conecta as regiões N- e C-terminal da toxina (domínio NC), e o outro é composto pelos resíduos 50-59 (domínio CT). Fragmentos de anticorpos específicos para a região próxima à ponte dissulfeto Cys12-Cys-63 inibiram a ligação de toxinas AaH II marcadas ao sítio do canal. Além disso, essas duas região eram inacessíveis aos anticorpos quando a toxina estava ligada ao canal.

A estrutura tridimensional da toxina AaH II mostrou que todos os resíduos importantes para a interação toxina-VSD estavam agrupados em uma face da toxina e sugeriam uma interação com múltiplos pontos do canal Nav. Os resíduos envolvidos na “região tóxica” pareciam pertencer às regiões C-terminal (domínio CT, resíduos 56-64) e N-terminal, que formam juntas o domínio NC. Também foi atribuída importância especial à alça de cinco

resíduos entre a fita β1 e a α-hélice (domínio RT) (Martin-Eauclaire et al., 2014).

As correlações encontradas apontam para uma determinada orientação da α-toxina ao interagir com o canal, a qual é condizente com a orientação proposta por Chugunov et al. em seu trabalho de 2013, que consiste no domínio NC apontando para o interior do canal (região do poro) e o domínio do núcleo apontando para a parte externa do canal (Chugunov et al., 2013). Com base em algumas das correlações encontradas no MOC- sistema 4 e informações da literatura, foi proposto um modo de interação de toxinas alfa com VSD-IV de canais Nav, mostrado na Figura 28.

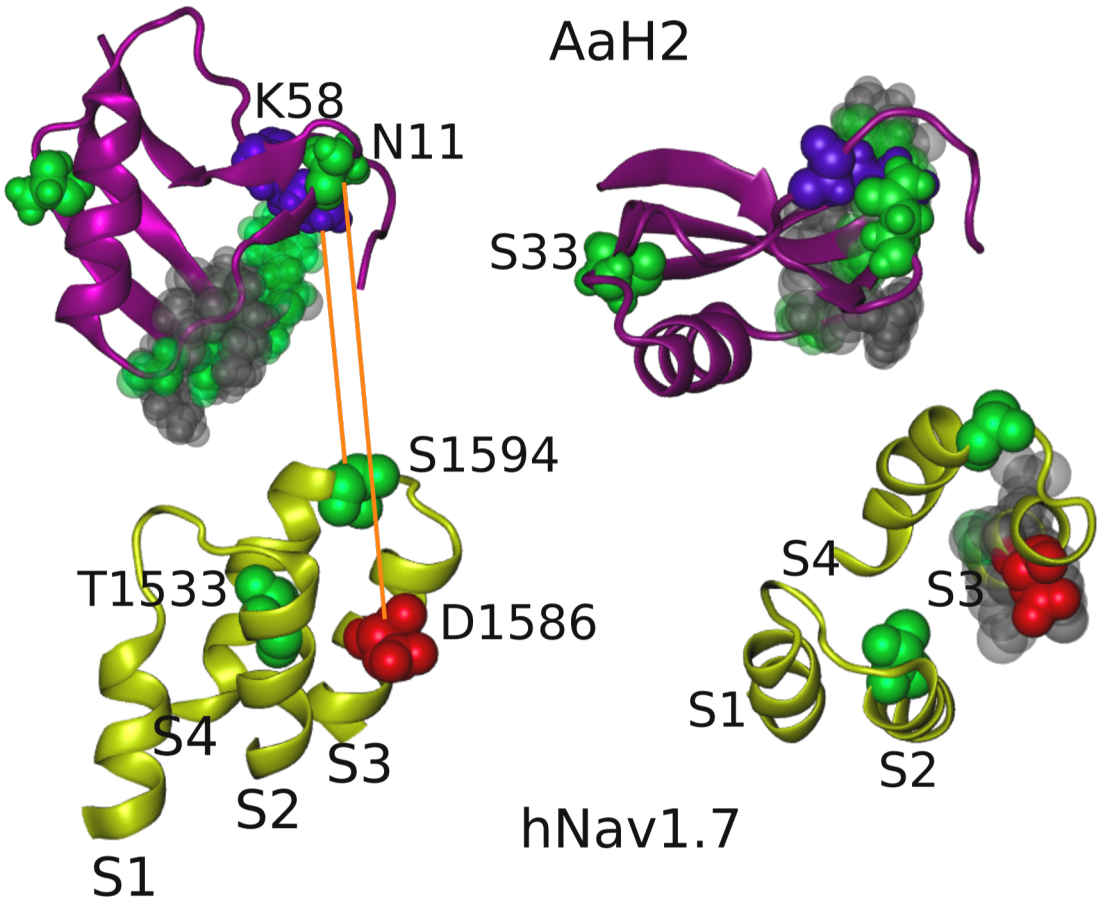

Figura 28: Possível modo de interação de uma toxina alfa com o sítio 3. A linha contínua foi traçada a partir de correlações obtidas da PCA do MOC – sistema 4 para o grupo α-toxinas–VSD-IV e indica que podem haver interações entre os grupos de resíduos interligados. À esquerda, são mostrados toxina e canal vistos de cima. É possível observar um cluster hidrofóbico na região inferior da toxina, formado pelos resíduos 15 a 17 e 38 a 44, que parece entrar em contato com um cluster hidrofóbico presente na região terminal da hélice S3. O resíduo S33 e T1533 são mostrados, pois também são importantes para a ligação da toxina com o canal. Na figura são mostrados a toxina AaH2 e o canal quimérico Na_VAb-$Na_V1.7$.

As correlações entre os CI 0 e CI 8 da toxina e os CI 0 e CI 8 (Glu1586 e Ser1594 em hNav1.7) do VSD-IV, respectivamente, ajudaram a determinar o posicionamento da toxina em relação ao canal. A orientação proposta possibilita a interação de um *cluster* hidrofóbico, localizado no domínio do núcleo de α-toxinas e essencial para sua função, com um *cluster* hidrofóbico na região S3 terminal do canal, comprovadamente necessário para a ligação da toxina.

Quando dois sítios de proteínas parceiras estão correlacionados, é provável que eles estejam em contato direto, entretanto, é possível que haja correlação entre sítios que não interagem diretamente. Em uma rede com muitas conexões, como a composta pelos resíduos de uma proteína dobrada, um grande conjunto de interações par a par pode gerar correlações estatísticas indiretas de longo alcance (Palovcak et al., 2014).

O fato de terem sido encontradas correlações entre um resíduo da toxina e vários resíduos do canal iônico significa que provavelmente há correlações indiretas entre alguns resíduos do canal e o resíduo da toxina e que o conjunto de resíduos do canal provavelmente está correlacionado por conta de sua proximidade estrutural. Como a metodologia utilizada não diferencia entre correlações diretas e indiretas, não é possível saber qual resíduo do canal está, de fato, interagindo com o resíduo na superfície da toxina.

Uma abordagem teórica que parece ser capaz de diferenciar sinais de acoplamento direto de sinais de acoplamento indireto é a análise de acoplamento direto (*direct-coupling analysis* – DCA). A DCA é capaz de definir resíduos importantes para a função da proteína, como resíduos presentes em sítios catalíticos e regiões de mudança conformacional (Hopf et al., 2012; Morcos et al., 2011; Weigt et al., 2009). Entretanto, a abordagem teórica utilizada neste trabalho já identifica resíduos importantes para a função das toxinas ao realizar a seleção dos canais de informação. Diante disso, a utilização da DCA provavelmente não traria contribuições significativas para a análise.

O trabalho de Waddell et al. deixa claro que evolução correlacionada é o que é detectado, enquanto coevolução é a causa hipotética (Waddell et al., 2007). Para a detecção de taxas de evolução correlacionada, deve-se utilizar tantos resíduos quanto possível para que também os sinais fracos sejam detectados. De fato, Kann et al. demonstram que é mais fácil detectar evolução correlacionada se mais sítios são considerados no cálculo (Kann et al., 2009).

Diante disso, uma forma de aprimorar os resultados obtidos seria através da análise de

múltiplos *hot-spots* de interação da toxina com o canal. Diversos dados experimentais indicam que as α- e β- toxinas apresentam mais de um ponto de interação com o canal. Em seu trabalho, Wang et al. identificaram como sítio de interação principal das α-toxinas alguns resíduos dos loops extracelulares S1-S2 e S3-S4 do domínio IV e como sítio de interação secundário resíduos do loop SS2-S6 do domínio I (Wang et al., 2011). Por outro lado, o sítio de ligação das β-toxinas não é bem definido.

Cestèle et al., usando a toxina Css4 de *Centruroides suffusus suffusus*, descobriram que essa neurotoxina se liga ao receptor nos *loops* extracelulares S3-S4 e S1-S2 do domínio II de canais $Na_V1.2$ (Cestèle et al., 1998, 2006); a toxina Css4 também se liga com menor afinidade ao *loop* S5-SS1 do domínio I e ao *loop* SS2-S6 do domínio III (Cestèle et al., 1998). Além disso, utilizando construções quiméricas do dos motivos de S3b-S4 de canais $rNa_V1.2$ e análise de ligação por espectroscopia de fluorescência, Campos et al. e Bosmans et al. descobriram que a toxina Ts1 de *Tityus serrulatus* não apenas se liga ao *loop* S3-S4 do domínio II mas também interage com o *loop* S3-S4 do domínio III, onde essa interação tem um efeito alostérico na ativação dos sensores de voltagem dos domínios I e IV de canais $Na_V1.2$ (Bosmans et al., 2008; Campos et al., 2007); Leipold et al., por sua vez, mostraram que a toxina Tz1 de *Tityus zulianus* se liga principalmente a região do poro SS2 no domínio III de canais $Na_V1.4$ (Leipold et al., 2006).

Banco de Estruturas

Ao longo da trajetória, os sistemas selecionados permaneceram estáveis, tendo em vista que todos os valores de RMSD se mantiveram abaixo de 3 Å, e que a variância foi pequena(Figura 23). Essa análise indica que as estruturas simuladas são conformacionalmente estáveis.

Em relação à superfície exposta, é possível notar que as médias de SASA obtidas são muito imprecisas, pois o desvio padrão é muito grande (Figura 24). Isso indica que existe uma variação significativa do parâmetro avaliado em cada uma das colunas do MSA, o que significa que o alinhamento de sequências primárias provavelmente não corresponde ao alinhamento de superfície das toxinas analisadas.

Conclusões

Interações do tipo proteína-proteína são essenciais para quase todos os processos celulares, responsáveis pelo funcionamento do organismo. O estudo dessas interações é, portanto, muito importante para um melhor entendimento dos sistemas moleculares. Nesse contexto, a abordagem teórica desenvolvida neste trabalho é uma contribuição inovadora, que facilita a visualização e o entendimento de sistemas em escala global, a partir de uma análise estatística do problema a ser investigado. A metodologia aqui proposta é capaz de, partindo de um conjunto de sequências primárias de polipeptídeos, as quais podem ser facilmente obtidas em bancos de dados, encontrar uma rede de coevolução que determina as melhores afinidades moleculares entre as proteínas analisadas. Além disso, é possível extrair do modelo otimizado de coevolução informações valiosas sobre quais seriam as regras moleculares que definem a rede de afinidades encontrada. A partir dessas regras, é possível identificar padrões de aminoácidos na superfície das proteínas analisadas que as caracterizem em termos funcionais. Uma vez identificados, esses padrões de aminoácidos podem, então, ser facilmente testados experimentalmente.

Perspectivas

Nos últimos 40 anos, houve um grande avanço na caracterização das peçonhas de escorpiões. Estudos mostram que, dentre os seus componentes, as neurotoxinas são os de maior importância médica, devido à sua alta toxicidade e relativa abundância (Gopalakrishnakone et al., 2015). Além disso, foi constatado que, ao se neutralizar os efeitos tóxicos das principais neurotoxinas, a toxicidade da peçonha é eliminada (Licea et al., 1996; Mendes et al., 2008). Nesse contexto, o estudo da interação de neurotoxinas e canais iônicos pode trazer diversas contribuições, principalmente para a área de saúde.

As neurotoxinas também podem ser ferramentas poderosas para o estudo de canais iônicos (Catterall et al., 2007; Cestèle and Catterall, 2000). O uso de toxinas no tratamento de insuficiência cardíaca, por exemplo, tem sido alvo de estudos por décadas, e a toxina de anêmona Anthopleurina-A indicou potencial de ser um agente farmacêutico eficiente com poucos efeitos colaterais (Hanck and Sheets, 1995). Caso um conhecimento mais aprofundado sobre a superfície funcional dessas toxinas seja adquirido, poderia tornar-se possível a engenharia de novas drogas para o tratamento de diversas condições patológicas associadas a canais iônicos.

Há também grande interesse na aplicação de neurotoxinas como agentes de controle biológico. Experimentos realizados demonstraram que substituições de resíduos em regiões específicas da toxina podem trazer alterações quanto a sua seletividade que levam à diminuição ou aumento da afinidade por determinado grupo de insetos (Moskowitz et al., 1994), o que pode ser interessante na produção de inseticidas seletivos. Além disso, há evidências de que o uso de neurotoxinas de escorpiões melhore a eficiência de inseticidas fúngicos (Harrison and Bonning, 2000; Wang and St Leger, 2007).

Além das contribuições no campo toxicológico, este trabalho traz importantes contribuições teóricas e metodológicas, pois foi desenvolvido utilizando uma abordagem nova que agrega conhecimento das áreas da matemática, estatística, biofísica e computação. A metodologia desenvolvida ainda precisa ser trabalhada e aprimorada, mas apresenta grande potencial de aplicação em diversos problemas que envolvem interações do tipo proteína-proteína. Os resultados obtidos das análises teóricas devem ajudar a diminuir os custos temporais e materiais necessários para a realização de trabalhos experimentais, servindo de guia para direcionar os experimentos.

Algumas questões foram levantadas durante a realização deste trabalho, sendo a principal delas referente a como as α- e β-toxinas de escorpião são capazes de manter uma estrutura 3D conservada e, ainda assim, possuir afinidade diferenciada por diversos sítios em diferentes tipos de canais Na_V. O banco estrutural de toxinas construído aqui pode ser utilizado para responder a essas questões, servindo como base de dados para um alinhamento múltiplo de aminoácidos localizados na superfície dessas moléculas. O objetivo é responder qual é o nível de acoplamento entre o núcleo e a superfície da toxina. Uma possível explicação para a questão proposta seria uma independência total ou parcial entre núcleo e superfície.

Para uma determinação mais precisa do modo de interação entre toxina e canal, será realizado *docking* molecular direcionado aliado a cálculos da energia livre de ligação. As estruturas utilizadas nas simulações serão de toxinas do banco estrutural gerado e de canais hNa_V modelados a partir das estruturas do canal Na_V bacteriano nos estados ativo e inativo. Além disso, alguns dos pares toxina-VSD do MOC serão selecionados para compor ensaios de mutagênese para confirmação experimental dos resultados.

Referências Bibliográficas

Ahern, C.A., Payandeh, J., Bosmans, F., and Chanda, B. (2016). The hitchhiker's guide to the voltage-gated sodium channel galaxy. J. Gen. Physiol. *147*, 1–24.

Bosmans, F., Martin-Eauclaire, M.-F., and Swartz, K.J. (2008). Deconstructing voltage sensor function and pharmacology in sodium channels. Nature *456*, 202–208.

Campos, F.V., Chanda, B., Beirão, P.S.L., and Bezanilla, F. (2007). beta-Scorpion toxin modifies gating transitions in all four voltage sensors of the sodium channel. J. Gen. Physiol. *130*, 257–268.

Catterall, W.A. (2000). From ionic currents to molecular mechanisms: The structure and function of voltage-gated sodium channels. Neuron *26*, 13–25.

Catterall, W.A., Cestèle, S., Yarov-Yarovoy, V., Yu, F.H., Konoki, K., and Scheuer, T. (2007). Voltage-gated ion channels and gating modifier toxins. Toxicon *49*, 124–141.

Cestèle, S., and Catterall, W.A. (2000). Molecular mechanisms of neurotoxin action on voltage-gated sodium channels. Biochimie *82*, 883–892.

Cestèle, S., Qu, Y., Rogers, J.C., Rochat, H., Scheuer, T., and Catterall, W.A. (1998). Voltage sensor-trapping: Enhanced activation of sodium channels by β- scorpion toxin bound to the S3-S4 loop in domain II. Neuron *21*, 919–931.

Cestèle, S., Yarov-Yarovoy, V., Qu, Y., Sampieri, F., Scheuer, T., and Catterall, W.A. (2006). Structure and Function of the Voltage Sensor of Sodium Channels Probed by a β-Scorpion Toxin. J. Biol. Chem. *281*, 21332–21344.

Cha, A., Ruben, P.C., George Jr., A.L., Fujimoto, E., and Bezanilla, F. (1999). Voltage Sensors in Domains III and IV, but Not I and II, Are Immobilized by Na+ Channel Fast Inactivation. Neuron *22*, 73–87.

Champeimont, R., Laine, E., Hu, S.-W., Penin, F., and Carbone, A. (2016). Coevolution analysis of Hepatitis C virus genome to identify the structural and functional dependency network of viral proteins. Sci. Rep. *6*, 26401.

Chen, R., and Chung, S.-H. (2012). Binding modes and functional surface of anti-mammalian scorpion α-toxins to sodium channels. Biochemistry (Mosc.) *51*, 7775–7782.

Chugunov, A.O., Koromyslova, A.D., Berkut, A.A., Peigneur, S., Tytgat, J., Polyansky, A.A., Pentkovsky, V.M., Vassilevski, A.A., Grishin, E.V., and Efremov, R.G. (2013). Modular Organization of α-Toxins from Scorpion Venom Mirrors Domain Structure of Their Targets, Sodium Channels. J. Biol. Chem. *288*, 19014–19027.

Cohen, L., Karbat, I., Gilles, N., Froy, O., Corzo, G., Angelovici, R., Gordon, D., and Gurevitz, M. (2004). Dissection of the Functional Surface of an Anti-insect Excitatory Toxin Illuminates a Putative "Hot Spot" Common to All Scorpion β-Toxins Affecting Na+ Channels. J. Biol. Chem. *279*, 8206–8211.

Cohen, L., Karbat, I., Gilles, N., Ilan, N., Benveniste, M., Gordon, D., and Gurevitz, M. (2005). Common Features in the Functional Surface of Scorpion β-Toxins and Elements That Confer Specificity for Insect and Mammalian Voltage-gated Sodium Channels. J. Biol. Chem. *280*, 5045–5053.

Couraud, F., Jover, E., Dubois, J.M., and Rochat, H. (1982). Two types of scorpion toxin receptor sites, one related to the activation, the other to the inactivation of the action potential sodium channel. Toxicon *20*, 9–16.

Cover, T.M., and Thomas, J.A. (2012). Elements of Information Theory (John Wiley & Sons).

Ehrlich, P.R., and Raven, P.H. (1964). Butterflies and Plants: A Study in Coevolution. Evolution *18*, 586–608.

Gopalakrishnakone, P., Possani, L.D., Schwartz, E.F., and Rodríguez de la Vega, R.C. (2015). Scorpion venoms.

Gordon, D., and Gurevitz, M. (2003). The selectivity of scorpion α-toxins for sodium channel subtypes is determined by subtle variations at the interacting surface. Toxicon *41*, 125–128.

Gordon, D., Karbat, I., Ilan, N., Cohen, L., Kahn, R., Gilles, N., Dong, K., Stühmer, W., Tytgat, J., and Gurevitz, M. (2007). The differential preference of scorpion α-toxins for insect or mammalian sodium channels: Implications for improved insect control. Toxicon *49*, 452–472.

Granier, C., Novotny, J., Fontecilla-Camps, J.C., Fourquet, P., el Ayeb, M., and Bahraoui, E. (1989). The antigenic structure of a scorpion toxin. Mol. Immunol. *26*, 503–513.

Gur, M., Kahn, R., Karbat, I., Regev, N., Wang, J., Catterall, W.A., Gordon, D., and Gurevitz, M. (2011). Elucidation of the molecular basis of selective recognition uncovers the interaction site for the core-domain of scorpion alpha-toxins on sodium channels. J. Biol. Chem. jbc.M111.259507.

Gurevitz, M., Gordon, D., Barzilai, M.G., Kahn, R., Cohen, L., Moran, Y., Zilberberg, N., Froy, O., Altman-Gueta, H., Turkov, M., et al. (2015). Molecular Description of Scorpion Toxin Interaction with Voltage-Gated Sodium Channels. In Scorpion Venoms, P. Gopalakrishnakone, L.D. Possani, E.F. Schwartz, and R.C.R. de la Vega, eds. (Springer Netherlands), pp. 471–491.

Hanck, D.A., and Sheets, M.F. (1995). Modification of inactivation in cardiac sodium channels: ionic current studies with Anthopleurin-A toxin. J. Gen. Physiol. *106*, 601–616.

Harrison, R.L., and Bonning, B.C. (2000). Use of Scorpion Neurotoxins to Improve the Insecticidal Activity of Rachiplusia ou Multicapsid Nucleopolyhedrovirus. Biol. Control *17*, 191–201.

Hassani, O., Mansuelle, P., Cestèle, S., Bourdeaux, M., Rochat, H., and Sampieri, F. (1999). Role of lysine and tryptophan residues in the biological activity of toxin VII (Ts γ) from the scorpion Tityus serrulatus. Eur. J. Biochem. *260*, 76–86.

Hille, B. (2001). Introduction. In *Ion channels of excitable membranes* (Vol. 507). Sunderland, MA: Sinauer, pp. 1–21.

Hopf, T.A., Colwell, L.J., Sheridan, R., Rost, B., Sander, C., and Marks, D.S. (2012). Three-dimensional structures of membrane proteins from genomic sequencing. Cell *149*, 1607–1621.

Huang, J., and MacKerell, A.D. (2013). CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. J. Comput. Chem. *34*, 2135–2145.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual molecular dynamics. J. Mol. Graph. *14*, 33–38.

Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics *11*, 431.

Kahn, R., Karbat, I., Ilan, N., Cohen, L., Sokolov, S., Catterall, W.A., Gordon, D., and Gurevitz, M. (2009). Molecular requirements for recognition of brain voltage-gated sodium channels by scorpion alpha-toxins. J. Biol. Chem. *284*, 20684–20691.

Kalia, J., Milescu, M., Salvatierra, J., Wagner, J., Klint, J.K., King, G.F., Olivera, B.M., and Bosmans, F. From Foe to Friend: Using Animal Toxins to Investigate Ion Channel Function. J. Mol. Biol.

Kann, M.G., Shoemaker, B.A., Panchenko, A.R., and Przytycka, T.M. (2009). Correlated Evolution of Interacting Proteins: Looking Behind the Mirrortree. J. Mol. Biol. *385*, 91–98.

Karbat, I., Frolow, F., Froy, O., Gilles, N., Cohen, L., Turkov, M., Gordon, D., and Gurevitz, M. (2004). Molecular Basis of the High Insecticidal Potency of Scorpion α-Toxins. J. Biol. Chem. *279*, 31679–31686.

Karbat, I., Turkov, M., Cohen, L., Kahn, R., Gordon, D., Gurevitz, M., and Frolow, F. (2007). X-ray Structure and Mutagenesis of the Scorpion Depressant Toxin LqhIT2 Reveals Key Determinants Crucial for Activity and Anti-Insect Selectivity. J. Mol. Biol. *366*, 586–601.

Kharrat, R., Darbon, H., Rochat, H., and Granier, C. (1989). Structure/activity relationships of scorpion α-toxins. Eur. J. Biochem. *181*, 381–390.

Lacroix, J.J., Campos, F.V., Frezza, L., and Bezanilla, F. (2013). Molecular Bases for the Asynchronous Activation of Sodium and Potassium Channels Required for Nerve Impulse Generation. Neuron *79*, 651–657.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinforma. Oxf. Engl. *23*, 2947–2948.

Legros, C., Céard, B., Vacher, H., Marchot, P., Bougis, P.E., and Martin-Eauclaire, M.-F. (2005). Expression of the standard scorpion alpha-toxin AaH II and AaH II mutants leading to the identification of some key bioactive elements. Biochim. Biophys. Acta BBA - Gen. Subj. *1723*, 91–99.

Leipold, E., Hansel, A., Borges, A., and Heinemann, S.H. (2006). Subtype Specificity of Scorpion β-Toxin Tz1 Interaction with Voltage-Gated Sodium Channels Is Determined by the Pore Loop of Domain 3. Mol. Pharmacol. *70*, 340–347.

Lewis, A.C.F., Saeed, R., and Deane, C.M. (2010). Predicting protein–protein interactions in the context of protein evolution. Mol BioSyst *6*, 55–64.

Licea, A.F., Becerril, B., and Possani, L.D. (1996). Fab fragments of the monoclonal antibody BCF2 are capable of neutralizing the whole soluble venom from the scorpion Centruroides noxius Hoffmann. Toxicon Off. J. Int. Soc. Toxinology *34*, 843–847.

Liu, L.-H., Bosmans, F., Maertens, C., Zhu, R.-H., Wang, D.-C., and Tytgat, J. (2005). Molecular basis of the mammalian potency of the scorpion α-like toxin, BmK M1. FASEB J. *19*, 594–596.

Lovell, S.C., and Robertson, D.L. (2010). An Integrated View of Molecular Coevolution in Protein–Protein Interactions. Mol. Biol. Evol. *27*, 2567–2575.

Madaoui, H., and Guerois, R. (2008). Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. Proc. Natl. Acad. Sci. U. S. A. *105*, 7708–7713.

Marcotte, P., Chen, L.-Q., Kallen, R.G., and Chahine, M. (1997). Effects of Tityus Serrulatus Scorpion Toxin gamma on Voltage-Gated Na sup + Channels. Circ. Res. *80*, 363–369.

Martin-Eauclaire, M.-F., Abbas, N., Céard, B., Rosso, J.-P., and Bougis, P.E. (2014). Androctonus Toxins Targeting Voltage-Gated Sodium Channels. In Scorpion Venoms, P. Gopalakrishnakone, E.F. Schwartz, L.D. Possani, and R.C.R. de la Vega, eds. (Springer Netherlands), pp. 1–25.

Mendes, T.M., Dias, F., Horta, C.C.R., Pena, I.F., Arantes, E.C., and Kalapothakis, E. (2008). Effective Tityus serrulatus anti-venom produced using the Ts1 component. Toxicon Off. J. Int. Soc. Toxinology *52*, 787–793.

Mintseris, J., and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein–protein interactions. Proc. Natl. Acad. Sci. U. S. A. *102*, 10930–10935.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. *108*, E1293–E1301.

Moskowitz, H., Herrmann, R., Zlotkin, E., and Gordon, D. (1994). Variability among insect sodium channels revealed by selective neurotoxins. Insect Biochem. Mol. Biol. *24*, 13–19.

Palovcak, E., Delemotte, L., Klein, M.L., and Carnevale, V. (2014). Evolutionary imprint of activation: The design principles of VSDs. J. Gen. Physiol. *143*, 145–156.

Pedraza Escalona, M., and Possani, L.D. (2013). Scorpion beta-toxins and voltage-gated sodium channels: interactions and effects. Front. Biosci. Landmark Ed. *18*, 572–587.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825−2830.

Peigneur, S., Cologna, C.T., Cremonez, C.M., Mille, B.G., Pucca, M.B., Cuypers, E., Arantes, E.C., and Tytgat, J. (2015). A gamut of undiscovered electrophysiological effects produced by Tityus serrulatus toxin 1 on NaV-type isoforms. Neuropharmacology *95*, 269–277.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kalé, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. J. Comput. Chem. *26*, 1781–1802.

Possani, L.D., Becerril, B., Delepierre, M., and Tytgat, J. (1999). Scorpion toxins specific for Na+-channels. Eur. J. Biochem. *264*, 287–300.

Quintero-Hernández, V., Jiménez-Vargas, J.M., Gurrola, G.B., Valdivia, H.H., and Possani, L.D. (2013). Scorpion venom components that affect ion-channels function. Toxicon *76*, 328–342.

Rogers, J.C., Qu, Y., Tanada, T.N., Scheuer, T., and Catterall, W.A. (1996). Molecular Determinants of High Affinity Binding of α-Scorpion Toxin and Sea Anemone Toxin in the S3-S4 Extracellular Loop in Domain IV of the Na+ Channel α Subunit. J. Biol. Chem. *271*, 15950–15962.

Schiavon, E., Pedraza-Escalona, M., Gurrola, G.B., Olamendi-Portugal, T., Corzo, G., Wanke, E., and Possani, L.D. (2012). Negative-shift activation, current reduction and resurgent currents induced by β-toxins from Centruroides scorpions in sodium channels. Toxicon *59*, 283–293.

Shannon, C., and Weaver, W. (1949). The Mathematical Theory of Information (University of Illinois Press).

Stock, L., Souza, C., and Treptow, W. (2013). Structural Basis for Activation of Voltage-Gated Cation Channels. Biochemistry (Mosc.) *52*, 1501–1513.

Swartz, K.J. (2008). Sensing voltage across lipid membranes. Nature *456*, 891–897.

Valen, L.V. (1973). Molecular evolution as predicted by natural selection. J. Mol. Evol. *3*, 89–101.

Vargas, E., Yarov-Yarovoy, V., Khalili-Araghi, F., Catterall, W.A., Klein, M.L., Tarek, M., Lindahl, E., Schulten, K., Perozo, E., Bezanilla, F., et al. (2012). An emerging consensus on voltage-dependent gating from computational modeling and molecular dynamics simulations. J. Gen. Physiol. *140*, 587–594.

Waddell, P.J., Kishino, H., and Ota, R. (2007). Phylogenetic Methodology for Detecting Protein Interactions. Mol. Biol. Evol. *24*, 650–659.

Wang, C., and St Leger, R.J. (2007). A scorpion neurotoxin increases the potency of a fungal insecticide. Nat. Biotechnol. *25*, 1455–1456.

Wang, J., Yarov-Yarovoy, V., Kahn, R., Gordon, D., Gurevitz, M., Scheuer, T., and Catterall, W.A. (2011). Mapping the receptor site for alpha-scorpion toxins on a Na+ channel voltage sensor. Proc. Natl. Acad. Sci. U. S. A. *108*, 15426–15431.

Webb, B., and Sali, A. (2014). Protein Structure Modeling with MODELLER. In Protein Structure Prediction, D. Kihara, ed. (Springer New York), pp. 1–15.

Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., and Hwa, T. (2009). Identification of direct residue contacts in protein–protein interaction by message passing. Proc. Natl. Acad. Sci. *106*, 67–72.

Whitley, D. (1994). A genetic algorithm tutorial. Stat. Comput. *4*, 65–85.

Yu, F.H., and Catterall, W.A. (2004). The VGL-Chanome: A Protein Superfamily Specialized for Electrical Signaling and Ionic Homeostasis. Sci STKE *2004*, re15.

Anexos

Anexo 1. *MSA seed de α- e β-NaScTxs*

```
SCX1_MESMA:    VRDAYIAKP-HNCVYECAR-----NEYCNDLCTKN--GAKSGYCQWVGKYGNGCWCIELPDNVPIRVPG------KCHR------
SCM2_MESMA:    VRDAYIAKP-HNCVYECAR-----NEYCNNLCTKN--GAKSGYCQWSGKYGNGCWCIELPDNVPIRVPG------KCH-------
SCX7_MESMA:    VRDGYIALP-HNCAYGCLN-----NEYCNNLCTKD--GAKIGYCNIVGKYGNACWCIQLPDNVPIRVPG------RCHPA-----
SCXN5_MESEU:   ARDAYIAKP-HNCVYECFDAF---SSYCNGVCTKN--GAKSGYCQILGTYGNGCWCIALPDNVPIRIPG------KCH-------
SCX3_LEIQU:    VRDAYIAKN-YNCVYECFR-----DSYCNDLCTKN--GASSGYCQWAGKYGNACWCYALPDNVPIRVPG------KCH-------
SCXA_LEIQH:    VRDAYIAKN-YNCVYECFR-----DAYCNELCTKN--GASSGYCQWAGKYGNACWCYALPDNVPIRVPG------KCHRK-----
SC12_MESMA:    VRDAYIAQN-YNCVYHCAR-----DAYCNELCTKN--GAKSGSCPYLGEHKFACYCKDLPDNVPIRVPG------KCHRR-----
SCXA_MESMA:    VRDAYIAKP-ENCVYECGI-----TQDCNKLCTEN--GAESGYCQWGGKYGNACWCIKLPDSVPIRVPG------KCQR------
SCX4_MESMA:    VRDAYIAKP-ENCVYHCAG-----NEGCNKLCTDN--GAESGYCQWGGRYGNACWCIKLPDDVPIRVPG------KCH-------
SCX1_ODODO:    VRDAYIADD-KNCVYTCAS-----NGYCNTECTKN--GAESGYCQWIGRYGNACWCIKLPDEVPIRIPG------KCR-------
SCX2_ANDAU:    VKDGYIVDD-VNCTYFCGR-----NAYCNEECTKL--KGESGYCQWASPYGNACYCYKLPDHVRTKGPG------RCHGR-----
SC11_MESMA:    VKDGYIADD-RNCPYFCGR-----NAYCDGECKKN--RAESGYCQWASKYGNACWCYKLPDDARIMKPG------RCNGG-----
SCX8_MESMA:    GRDAYIADS-ENCTYFCGS-----NPYCNDVCTEN--GAKSGYCQWAGRYGNACYCIDLPASERIKEPG------KCG-------
SCL3_LEIQH:    VRDGYIAQP-ENCVYHCFPG----SSGCDTLCKEK--GGTSGHCGFKVGHGLACWCNALPDNVGIIVEG-----EKCHS------
SCX1_MESTA:    GEDGYIADG-DNCTYICTF-----NNYCHALCTDK--KGDSGACDWWVPYGVVCWCEDLPTPVPIRGSG------KCR-------
SCXV_CENSC:    KKDGYPVDS-GNCKYECLKD----D-YCNDLCLER--KADKGYCYWG---KVSCYCYGLPDNSPTKTSG------KCNPA-----
KURT_PARTR:    KIDGYPVDY-WNCKRICWYN----NKYCNDLCKGL--KADSGYCWGW---TLSCYCQGLPDNARIKRSG------RCRA------
SIX1_MESMA:    KKNGYAVDS-SGKVSEC-----LLNNYCNNICTKVY-YATSGYCCLL-----SCYCFGLDDDKAVLKIKDATKS-YCDVQIIG--
SIXP_MESMA:    KKNGYAVDS-SGKVAEC-----LFNNYCNNECTKVY-YADKGYCCLL-----KCYCFGLADDKPVLDIWDSTKN-YCDVQIIDLS
SIXE_HOTJU:    KKNGYPLDR-NGKTTECSGVNAIAPHYCNSECTKVY-YAESGYCCWG-----ACYCFGLEDDKPIGPMKDITKK-YCDVQIIPS-
SCXI_CENSC:    AKDGYLVEK-TGCKKTCYK-LGEND-FCNRECKWKHIGGSYGYCYGF-----GCYCEGLPDSTQTWPL--PNK--TCGKK-----
Q6V4Z0_CENSC: -KDGYLVEK-TGCKKTCYK-LGEND-FCNRECKWKHIGGSYGYCYGF-----GCYCEGLPDSTQTWPL--PNK--TC--------
SCX2_CENSC:    AKEGYLVNKSTGCKYGCLK-LGENE-GCDKECKAKNQGGSYGYCYAF-----ACWCEGLPESTPTYPL--PNK--SCSRK-----
SCX3_CENSC:    AKEGYLVNKSTGCKYGCLK-LGENE-GCDKECKAKNQGGSYGYCYAF-----ACWCEGLPESTPTYPL--PNK--SCGKK-----
SCX5_CENNO:    AKEGYLVNKSTGCKYGCLL-LGKNE-GCDKECKAKNQGGSYGYCYAF-----GCWCEGLPESTPTYPL--PNK--SCSKK-----
SCX1_CENSC:    AKEGYLVKKSDGCKYDCFW-LGKNE-HCDTECKAKNQGGSYGYCYAF-----ACWCEGLPESTPTYPL--PNK--SCGKK-----
SCX2_CENNO:    AKEGYLVDKNTGCKYECLK-LGDND-YCLRECKQQYGKGAGGYCYAF-----ACWCTHLYEQAIVWPL--PNK--RCSGK-----
SCX2_CENSU:    -KEGYLVSKSTGCKYECLK-LGDND-YCLRECKQQYGKSSGGYCYAF-----ACWCTHLYEQAVVWPL--PNK--TCN-------
SCX12_CENNO:   -RDGYPLAS-NGCKFGCSG-LGENNPTCNHVCEKK-AGSDYGYCYAW-----TCYCEHVAEGTVLWGD--SGTG-PCRS------
SCX1_TITSE:    CKEGYLMDH-EGCKLSC---FIRPSGYCGRECGIK--KGSSGYCAWP-----ACYCYGLPNWVKVWDR--ATN--KCGKK-----
SCX5_CENSC:    -KDGYPVDS-KGCKLSC---VAN--NYCDNQCKMK--KASGGHCYAM-----SCYCEGLPENAKVSDS--ATN--IC--------
SIX2_LEIQH:    NADGYIKRR-DGCKVACL--IGNEG--CDKECKAY--GGSYGYCWTWG---LACWCEGLPD-DKTWKS--ETN--TCGGKK----
```

Anexo 2. *Algoritmo em Python para cálculo da entropia (H)*

```
"""

This algorithm computes entropy (H) in bits for each position of a multiple
sequence alignment (MSA). A reduced amino acids alphabet, as well as the
parameters theta (similarity cutoff) and lambda (pseudocounter) are considered
here to improve statistics.

Inspired in code by: Werner Treptow
Code by: Caio Souza and Camila Pontes

The code was adapted to work using matrix arithmetics from NumPy.

"""
import numpy as np
from scipy import spatial
from Bio import AlignIO

# USER DEFINED
MSA = "msa.fasta"           # INPUT: MSA file name
Q = 5                       # Number of possible letters after encoding the MSA
THETA = 0.95                # Sequence similarity cutoff
LAMBDA = 0.5                # Pseudo-counter
OUTPUT = "entropy.out"      # OUTPUT: output file name

# READ MSA
msa = AlignIO.read(MSA, "fasta")
seqlen = msa.get_alignment_length()  # Number of sequences
seqnum = len(msa)                    # Number of aligned positions

# ENCODE MSA:       0 - hydrophobic (A, V, I, L, M, F, W, Y, P, C, G);
            #   1 - positively charged (R, H, K);
            #   2 - negatively charged (D, E);
            #   3 - not charged (S, T, N, Q);
            #   4 - gap
aminos = {"A": 0, "R": 1, "N": 3, "D": 2, "Q": 3,
          "E": 2, "G": 0, "H": 1, "L": 0, "K": 1,
          "M": 0, "F": 0, "S": 3, "T": 3, "W": 0,
          "Y": 0, "C": 0, "I": 0, "P": 0, "V": 0,
          "-": 4, "X": 4, "B": 4}

encoded_msa = np.empty((seqnum, seqlen), dtype="int")
for (i, j), A in np.ndenumerate(msa):
    encoded_msa[i,j] = aminos[A.upper()]

# WEIGHT SEQUENCES
hammdist = spatial.distance.pdist(encoded_msa, 'hamming')
weight_matrix = spatial.distance.squareform(hammdist < (1.0 - THETA))
weight = 1.0 / (np.sum(weight_matrix, axis=1) + 1.0)
# Calculate and print the effective number of sequences (Meff)
Meff = np.sum(weight)
```

```
print("The effective number of sequences is Meff = {0}".format(Meff))

# CALCULATES SITE FREQUENCIES
aa = np.arange(Q)
sitefreq = np.sum(np.multiply(weight[np.newaxis,:,np.newaxis],
encoded_msa[np.newaxis,:,:] == aa[:,np.newaxis,np.newaxis]), axis=1)
# Formula (4)
sitefreq = (LAMBDA / (LAMBDA*Q + Meff*Q)) + sitefreq / (LAMBDA+Meff)

# CALCULATE ENTROPY: Formula (1)
entropy = -np.nansum(sitefreq * np.log2(sitefreq), axis=0)

# OUTPUT DATA
np.savetxt(OUTPUT, entropy)
```

Anexo 3. *Algoritmo em Python para o cálculo da informação mútua (MI)*

```
"""

This algorithm computes conditional entropy using the formula H(X|Y) = H(X,Y) -
H(Y), as well as mutual information using MI(X;Y) = H(X) - H(X|Y), in bits, for
each position of a multiple sequence alignment (MSA) of NaScTxs, between the
variables X: 'amino acids type' and Y: 'toxin type'. A reduced amino acids
alphabet and the parameters theta (similarity cutoff) and lambda (pseudocounter)
are considered here to improve statistics.

Inspired in code by: Werner Treptow
Code by: Caio Souza and Camila Pontes

The code was adapted to work using matrix arithmetics from NumPy.

"""
import numpy as np
from scipy import spatial
from Bio import AlignIO

def cantor(x, y):
    """ Cantor pairing function.
    This function encodes a tuple of two natural integers into a single integer.
"""
    return (x + y) * (x + y + 1) / 2 + y

# USER DEFINED
MSA = "msa.fasta"           # INPUT: MSA file name
Q = 5                       # Number of possible letters after encoding the MSA
THETA = 0.95                # Sequence similarity cutoff
LAMBDA = 0.5                # Pseudo-counter
OUTPUT = "entropy.out"      # OUTPUT: conditional entropy output file name
BREAK = 175                 # First alpha-toxin ID

# READ MSA
msa = AlignIO.read(MSA, "fasta")
seqlen = msa.get_alignment_length()
seqnum = len(msa)

# DEFINE MSA SUBSETS (from ID-0 to ID-174: beta-toxins; from ID-175 to ID-289:
alpha-toxins)
subMask = np.zeros((seqnum))
subMask[0:BREAK] = 1

# ENCODE MSA:       0 - hydrophobic (A, V, I, L, M, F, W, Y, P, C, G);
            #   1 - positively charged (R, H, K);
            #   2 - negatively charged (D, E);
            #   3 - not charged (S, T, N, Q);
            #   4 - gap
aminos = {"A": 0, "R": 1, "N": 3, "D": 2, "Q": 3,
          "E": 2, "G": 0, "H": 1, "L": 0, "K": 1,
          "M": 0, "F": 0, "S": 3, "T": 3, "W": 0,
```

```
            "Y": 0, "C": 0, "I": 0, "P": 0, "V": 0,
            "-": 4, "X": 4, "B": 4}

encoded_msa = np.empty((seqnum, seqlen), dtype="int")
for (i, j), A in np.ndenumerate(msa):
    encoded_msa[i,j] = aminos[A.upper()]

# WEIGHT SEQUENCES
hammdist = spatial.distance.pdist(encoded_msa, 'hamming')
weight_matrix = spatial.distance.squareform(hammdist<(1.0 - THETA))
weight = 1.0/(np.sum(weight_matrix, axis=1)+1.0)
# Calculate and print the effective number of sequences (Meff)
Meff=np.sum(weight)
print("The effective number of sequences is Meff = {0}".format(Meff))

# CALCULATE H(Y)
asubProb = np.sum(np.compress(subMask == 0, weight))
bsubProb = np.sum(np.compress(subMask == 1, weight))
# Formula (4)
asubProb = (LAMBDA / (LAMBDA*2 + Meff*2)) + asubProb / (LAMBDA+Meff)
# Formula (4)
bsubProb = (LAMBDA / (LAMBDA*2 + Meff*2)) + bsubProb / (LAMBDA+Meff)
# Formula (1)
yEnt = - asubProb*np.log2(asubProb) - bsubProb*np.log2(bsubProb)

# CALCULATE H(X,Y)
pairEnt = np.zeros((seqlen,), dtype="float64")
probMask = np.empty(seqnum, dtype=np.bool_)
for i in range(seqlen):
    c = cantor(encoded_msa[:,i], subMask)
    unique, aaIdx = np.unique(c, True)
    pairProb = np.sum((np.multiply(weight,(c - c[:,np.newaxis]) == 0)), axis=1)
    # Formula (5)
    pairProb = (LAMBDA / (LAMBDA*Q*2 + Meff*Q*2)) + pairProb / (LAMBDA + Meff)
    probMask[:] = True
    probMask[aaIdx] = False
    pairProb[probMask] = 0
    p0 = (LAMBDA / (LAMBDA*Q*2 + Meff*Q*2))
    pairEnt[i] = -np.nansum(pairProb * np.log2(pairProb))
    if (LAMBDA != 0):
        pairEnt[i] -= (Q*2-len(unique))*p0*np.log2(p0)

# CALCULATE H(X|Y)
condEnt = pairEnt - yEnt                              # H(X|Y) = H(X,Y) - H(Y)

# OUTPUT DATA
np.savetxt(OUTPUT, condEnt)                           # Save H(X|Y)
entAB = np.loadtxt("entropy_input_file")              # Load H(X)
np.savetxt("mi_output_file", entAB - condEnt)         # Save MI(X;Y)
```

Anexo 4. *Lista de identificadores do UniProtKB de todas as α- e β-toxinas simuladas que fazem parte do banco estrutural, totalizando 14 α- e 9 β-toxinas do seed (com cristal disponível no PDB) e 36 α- e 23 β-toxinas modeladas, totalizando 82 toxinas*

ID	Toxina	Tipo	Molde
1	SCX1_ODODO	α	-
2	SCXA_LEIQH	α	-
3	SC11_MESMA	α	-
4	SC12_MESMA	α	-
5	SCX8_MESMA	α	-
6	SCX2_ANDAU	α	-
7	SCXA_MESMA	α	-
8	SCX3_LEIQU	α	-
9	SCXV_CENSC	α	-
10	SCX4_MESMA	α	-
11	SCM2_MESMA	α	-
12	SCX1_MESMA	α	-
13	SCX5_CENSC	α	-
14	SCX7_MESMA	α	-
15	MKTX3_MESMA	α	SCM2_MESMA
16	SC10_MESMA	α	SC11_MESMA
17	SC13_MESMA	α	SC12_MESMA
18	KURT1_PARGR	α	KURT_PARTR
19	SC15_MESMA	α	SCXN5_MESEU
20	MKTX2_MESMA	α	SCX1_MESMA
21	SC17_MESMA	α	SCXN5_MESEU
22	SCA1_MESMA	α	SCXN5_MESEU
23	SCA4_MESMA	α	SCXN5_MESEU
24	SCAA_MESMA	α	SCXN5_MESEU
25	SCIT_HOTJU	α	SCX8_MESMA
26	SCL7_LEIQH	α	SC12_MESMA
27	SCU1_MESMA	α	SCXN5_MESEU
28	SCX1_BUTMA	α	SCX1_ODODO
29	SCX28_HOTTS	α	SCX1_ODODO
30	SCX2_ANDAM	α	SCX8_MESMA
31	SCX2_LEIQH	α	SCXN5_MESEU

32	SCX3_BUTOC	α	SCXN5_MESEU
33	SCX3_BUTOM	α	SCXN5_MESEU
34	SCX3_MESMA	α	SCXN5_MESEU
35	SCX3_ORTSC	α	SCXN5_MESEU
36	SCX47_MESMA	α	SCM2_MESMA
37	SCX4_LEIQH	α	SCX1_ODODO
38	SCX4_LEIQU	α	SCX1_ODODO
39	SCX5_LEIQU	α	SCXN5_MESEU
40	SCX8_ANDMA	α	SCXN5_MESEU
41	SCXA_ANDAU	α	SCXN5_MESEU
42	SCXA_BUTOM	α	SCXN5_MESEU
43	SCXB_ANDAU	α	SCX8_MESMA
44	SCXB_BUTOC	α	SCXN5_MESEU
45	SCXB_BUTOM	α	SCXN5_MESEU
46	SCXB_MESMA	α	SCXN5_MESEU
47	SCXC_BUTOM	α	SCM2_MESMA
48	SCXD_BUTOM	α	SC12_MESMA
49	SIX1_BUTOC	α	SCM2_MESMA
50	SCXN2_MESEU	α	SCXN5_MESEU
51	Q6V4Z0_CENSC	β	-
52	SCX1_TITSE	β	-
53	SCX2_CENSC	β	-
54	SCX3_CENSC	β	-
55	SCXI_CENSC	β	-
56	SCX1_CENSC	β	-
57	SCX5_CENNO	β	-
58	SIX1_MESMA	β	-
59	SIXE_HOTJU	β	-
60	AEP1_MESMA	β	SIX2_LEIQH
61	SCNA7_TITDI	β	SCX1_TITSE
62	SCX10_CENEX	β	SCX2_CENNO
63	SCX12_TITOB	β	SCX1_TITSE
64	SCX1_CENNO	β	SCX2_CENNO
65	SCX1_TITBA	β	SCX1_TITSE

66	SCX1_TITST	β	SCX1_TITSE
67	SCX1_TITTR	β	SCX1_TITSE
68	SCX39_CENSU	β	SCX5_CENNO
69	SCX4_CENEX	β	Q6V4Z0_CENSC
70	SCX6_CENEX	β	Q6V4Z0_CENSC
71	SCX7_CENEX	β	Q6V4Z0_CENSC
72	SCX7_TITCO	β	SCX1_TITSE
73	SCX9_CENEX	β	SCX2_CENNO
74	SCXR_CENLL	β	Q6V4Z0_CENSC
75	SCXX_CENNO	β	SCX5_CENNO
76	SIX2_BUTAR	β	SIX2_LEIQH
77	SIX2_LEIQU	β	SIX2_LEIQH
78	SIX3_MESMA	β	SIX2_LEIQH
79	SIX4_BUTOC	β	SIX2_LEIQH
80	SIX5_BUTOC	β	SIX2_LEIQH
81	SIXI_ORTSC	β	SIX2_LEIQH
82	SX25_LEIQH	β	SIX2_LEIQH