28. Ciência, Tecnologia e Inovação

Modelagem e decomposição de redes de coevolução de aminoácidos: Aplicações na determinação de especificidade e anotação de proteínas

Fonseca Jr, Neli; nelijfjr@gmail.com; Carrijo, Lucas; lucas.carrijodeoliveira@gmail.com;
Querino, Marcelo; marceloqla@gmail.com
Universidade Federal de Minas Gerais

Resumo

Estudos de evolução molecular por técnicas computacionais são geralmente conduzidos a partir de um alinhamento múltiplo de sequências homólogas, no qual sequências provavelmente originadas por um ancestral comum são alinhadas de forma que aminoacidos equivalentes ocupem a mesma posição. Padrões de conservação de resíduos em um alinhamento, ou em um subconjunto de suas sequências, podem ser extremamente informativos por sugerirem posições sob seleção e restrição evolutiva. Neste trabalho desenvolvemos uma metodologia baseado em ciências das redes, com objetivo de identificar grupos de resíduos e propriedades estereoquímicas funcionalmente relacionados, através da análise de coocorrência e detecção de comunidades. A metodologia foi aplicada as famílias das Transtiretinas/HIUases e dos receptores acoplados a proteína G (classe A). Em ambos os casos foram obtidos com sucesso grupos de resíduos determinantes de especificidade para diversas subclasses funcionais. Estes dados foram posteriormente utilizados como estimadores para uma máquina de suporte de vetores (SVM) que foi capaz de classificar corretamente até mesmo subclasses, a quais nenhum resíduo específico foi identificado. A classificação por SVM foi também aplicada as GPCRs órfãs gerando novas hipóteses a respeito das classes funcionais destas sequências.

Palavras chave: Bioinformática Evolutiva, Bioinformática Funcional, Ciências das Redes, Aprendizagem de Máquina

28. Ciência, Tecnologia e Inovação

Modelagem e decomposição de redes de coevolução de aminoácidos: Aplicações na determinação de especificidade e anotação de proteínas

Fonseca Jr, Neli; Carrijo, Lucas; Querino, Marcelo
nelijfjr@gmail.com; lucas.carrijodeoliveira@gmail.com; marceloqla@gmail.com
Programa de Pós-Graduação em Bioinformática
Universidade Federal de Minas Gerais

Resumo

Estudos de evolução molecular por técnicas computacionais são geralmente conduzidos a partir de um alinhamento múltiplo de sequências homólogas, no qual sequências provavelmente originadas por um ancestral comum são alinhadas de forma que aminoacidos equivalentes ocupem a mesma posição. Padrões de conservação de resíduos em um alinhamento, ou em um subconjunto de suas sequências, podem ser extremamente informativos por sugerirem posições sob seleção e restrição evolutiva. Neste trabalho desenvolvemos uma metodologia baseado em ciências das redes, com objetivo de identificar grupos de resíduos e propriedades estereoquímicas funcionalmente relacionados, através da análise de coocorrência e detecção de comunidades. A metodologia foi aplicada as famílias das Transtiretinas/HIUases e dos receptores acoplados a proteína G (classe A). Em ambos os casos foram obtidos com sucesso grupos de resíduos determinantes de especificidade para diversas subclasses funcionais. Estes dados foram posteriormente utilizados como estimadores para uma máquina de suporte de vetores (SVM) que foi capaz de classificar corretamente até mesmo subclasses, a quais nenhum resíduo específico foi identificado. A classificação por SVM foi também aplicada as GPCRs órfãs gerando novas hipóteses a respeito das classes funcionais destas sequências.

Palavras chave: Bioinformática Evolutiva, Bioinformática Funcional, Ciências das Redes, Aprendizagem de Máquina

Introdução

Estudos computacionais envolvendo dados biológicos tem sido realizados desde a década de 50, quando Bennet e Kendrew publicaram um programa de computador utilizado na determinação da primeira estrutura em alta resolução de uma proteína, a mioglobina [Bennett & Kendrew, 1952; Kendrew et al., 1958]. Já na década de 60, Linus Paulling e Emile Zuckerland, através de estudos com um pequeno conjunto de sequências de hemoglobina, perceberam que as sequências biológicas evoluem com taxas mensuráveis e relativamente constantes [Zuckerkandl & Pauling, 1962, 1965]. Pouco tempo depois, Kimura et al. propôs a teoria neutra da evolução molecular,

segundo a qual, em nível molecular, a maior parte da variabilidade genética nas espécies é seletivamente neutra, ou seja, a maioria dos aminoácidos de uma proteína podem passar por mutações aleatórias sem nenhuma alteração em sua função, estando apenas alguns poucos sítios sub uma restrição evolutiva mais rigorosa [Kimura et al., 1968]. Estes trabalhos foram fundamentais para o surgimento e avanço da biologia molecular evolutiva, pois possibilitaram que padrões evolutivos pudessem ser extraídos de um conjunto de sequências homólogas, levando a informações a respeito da função, estrutura e história biológica das mesmas.

Uma das principais formas de extrair padrões evolutivos de sequências se da através do uso de alinhamentos múltiplos de sequências (AMS). Um AMS consiste de um conjunto de sequências de proteína ou DNA alinhadas de forma que suas respectivas posições se mantenham nas mesmas colunas. Caso estas sequências possuam uma relação de homologia, é possível dizer que os padrões variabilidade de aminoácidos em cada coluna representam uma manifestação de substituições impostas pela função [Dima & Thirumalai, 2006]. Portanto aminoácidos estritamente conservados em um AMS são geralmentes utilizados como preditores de importância funciona ou estrutural [Choi et al., 2012; Pazos & Bang, 2006]. Padrões de conservação também podem ser observados em relação a propriedades físico-químicas ou estruturais necessárias para que a proteína conserve sua atividade ou estabilidade [Chakrabarti et al., 2007].

Ao analisar os padrões de conservação em um AMS de forma local, pode-se observar os sítios determinantes de especificidade. Este tipo de padrão constitui-se de grupos extremamente conservados em uma subfamília, porém variável nas outras. Chakraborty & Chakrabarti [2014] definiram dois tipos de sítios determinantes de especificidade. O tipo 1 se refere a casos, no qual os grupos determinantes de cada subfamília possuem restrições evolutivas completamente distintas, geralmente estando associado a especificidade funcional. O tipo 2 ocorre quando a posição é conservada em mais de uma subfamília, porém o aminoácido ou propriedade que define a especificidade varia conforme cada subclasse. Este tipo de padrão é mais frequente em enzimas, estando geralmente atrelado a especificidade em relação ao ligante [Chakraborty & Chakrabarti, 2014].

A detecção de sítios determinantes de especificidade pode ser aplicada em diversos contextos, como: estudos de evolução molecular, uma vez que permite rastrear eventos como neofuncionalização ou mesmo auxiliar na reconstrução de sequências ancestrais; estudos de

caracterização de famílias de proteínas, visto que tais sítios estão geralmente associados a uma importância estrutural ou funcional específica de cada subclasse; anotação de sequências, por se tratarem de ótimos estimadores; e trabalhos de mutagêneses de sítio dirigida, pois estes resíduos são fortes candidatos a alvo de inibição [Bleicher et al., 2011; da Fonseca Jr et al., 2017; Afonso et al., 2013; Pedruzzi et al., 2014; Rios-Anjos et al., 2017; Suhadolnik et al., 2017].

Um dos principais problemas na detecção de determinantes de especificidade é a escalaridade. Em uma recente revisão acerca dos métodos disponíveis na literatura, Chakraborty & Chakrabarti [2014] observaram que apenas um terço dos métodos avaliados eram computacionalmente plausíveis de serem utilizados em larga escala. Além disto, os testes foram realizados em alinhamentos com números extremamente reduzido de sequências. O maior AMS utilizado possuía 180 sequências, quantidade ínfima se comparado aos alinhamentos disponíveis no Pfam, que podem passar de 1 milhão de sequências. Outro ponto é que muitos métodos acabam focando na determinação de posições determinantes de especificidade, portanto acabam por falhar na detecção dos determinantes de tipo 1. Um último problema comum é a necessidade de conhecimentos a priori, algorítimos que pedem informações extras como árvores filogenéticas, anotações ou subfamílias predefinidas. Este tipo de abordagem acaba sendo inviável para análises exploratórias e de caracterização

As anotações automática de novas proteínas é normalmente realizada através de análises de identidade global entre as sequências, o que levam os bancos de dados a possuírem taxas de erro de até 80% [Schnoes et al. 2009]. Este tipo de erro geralmente ocorre dentro das superfamílias. Logo, caso seja possível determinar grupos de resíduos que realmente determinem a especificidade de uma subfamília, seria plausível desenvolver um classificado capaz de realizar anotações automáticas de sequências com uma alta taxa de acerto dentro das superfamílias.

Objetivos

Este trabalho tem como objetivo o desenvolvimento de um novo algorítimo escalar e computacionalmente eficiente para a detecção de resíduos determinantes de especificidade, através da modelagem, validação e detecção de comunidades em redes de coocorrência de resíduos. Além disto, foi avaliado também a capacidade de classificar sequências corretamente utilizando apenas os dados obtidos na detecção dos determinantes de especificidade. A metodologia proposta neste trabalho está incluída no software PFstats [Fonseca-Júnior, Néli J., et al., 2018].

Materiais e Métodos

A primeira etapa da metodologia proposta neste trabalho consiste em remover possíveis vieses do AMS. Para isto, são aplicados quatro filtros: cobertura, identidade máxima entre sequências, frequência máxima e frequência mínima dos nós. A filtragem por cobertura consiste em remover possíveis fragmentos de sequências que possam estar presentes no alinhamento, assim, sequências que possuem uma baixa cobertura em relação as posições válidas do perfil HMM são descartadas. A remoção de sequências por identidade máxima tem o objetivo de remover possíveis vieses causado pelo acúmulo de sequências com alta identidade. As últimas duas etapas consistem em remover nós (resíduos ou propriedades) de acordo com sua frequência máxima e mínima no AMS. Os nós extremamente conservados são removidos com o objetivo de reduzir a complexidade da rede. Já a remoção pela frequência mínima é aplicada com o objetivo de remover possíveis ruídos. Nós com a frequência extremamente baixa podem gerar falsos positivos pela falta de amostragem suficiente.

A modelagem de rede parte da observação de um AMS como uma rede bipartida, no qual o conjunto U é formado pelos identificadores das sequências, e o conjunto V é formado por todos os possíveis resíduos (i.e. aminoácido seguido por sua posição no AMS). Desta forma, ao projetar a rede monopartida de V, grupos de resíduos que coocorrem nas mesmas sequências tendem a formar comunidades.

A rede pode ser ampliada para incluir os sinais de propriedades físico-químicas. Para isto, novos nós representando estas propriedades são adicionados. Ao incluir este tipo de padrão, é necessário filtrar os nós sinônimos, sendo assim, um algorítimo é aplicado com o intuito de manter em cada vizinhança um único nó correspondente a cada posição. O nó selecionado será aquele que possua a menor distância média em relação aos seus vizinhos (distância do cosseno calculada a partir da matriz de biadjacência). Caso mais de um nó compartilhem um mesmo valor de distância, será mantido o que representar um menor subconjunto.

A geração de projeções tende a produzir redes extremamente densas e não basta a simples aplicação de um threshold. De acordo com Neal, a aplicação de cortes simples em redes projetadas possui três principais deficiências: o viés de arbitrariedade, ou seja, a utilização de um valor simplesmente arbitrário no corte; o viés estrutural, Watts demonstrou que a aplicação de um threshold incondicional irá sempre produzir redes com alto coeficiente de agrupamento, não pelas características estruturais da rede, mas

por um viés gerado pela remoção de arestas; e finalmente, o viés de não escalaridade, uma que vez que os pesos das arestas da projeção são diretamente correlacionados ao seus respectivos graus na rede bipartida. Os grupos que coocorrem em proporções menores seriam simplesmente descartados.

Neste trabalho foi aplicado a abordagem de Tumminello et al. 2011 para validação e normalização das arestas. O método consiste em validar cada aresta de uma projeção monopartida através de uma hipótese nula de coocorrência aleatória de vizinhos em comuns, levando em consideração a heterogeneidade dos elementos de ambos os conjuntos da rede bipartida.

A grande maioria dos algorítimos de detecção de comunidades em redes tem como principio básico o particionamento em função da minimização da modularidade [Fortunato & Hric, 2016]. Uma vez que neste trabalho o foco principal é determinar grupos coocorrentes no AMS, independente da estrutura da rede, foi desenvolvido um algorítimo para detecção de comunidades em redes projetadas baseado em um clustering hierárquico aglomerativo.

Em seu estágio inicial, cada nó da rede é atribuído a uma comunidade própria. Nas etapas posteriores, é calculado a distância do cosseno entre cada par de comunidade. Os vetores representantes de cada comunidade são formados pela média das colunas de seus nós na matriz de biadjacência. Os pares de comunidade que apresentam a menor distância são fundidos. O algorítimo finaliza quando não há nenhum par de comunidades cuja distância é menor do que um dado valor (neste trabalho foi utilizado 0.4 como limiar).

Na etapa de classificação, cada sequência do alinhamento é representada por um vetor , de tamanho N, sendo este o número de comunidades detectadas nas redes. Cada posição v_i é composta pela média dos resíduos da comunidade i presentes na sequência. Estes dados são utilizados como características para alimentar uma máquina de suporte de vetores (SVM). A máquina é treinada apenas com sequências extraídas do Swiss-Prot, e por se tratar de um conjunto com um número geralmente limitado de sequências, a abordagem escolhida para o treinamento foi a LOOCV (Leave-one-out Cross Validation). Esta abordagem consiste em a cada passo, separar uma única sequência para a etapa de validação, enquanto todas as outras são utilizadas no treinamento. Este processo é repetido até que todas as sequências do conjunto de treinamento tenham sido utilizadas na validação.

Como forma de validação da metodologia foram conduzidos estudos de casos com a família dos receptores acoplados a proteína G (classe A) (Pfam: PF00001) e com a família das HIUases e Transtiretinas (Pfam: PF00576). Ambos os alinhamentos foram obtidos a partir do Pfam e continham respectivamente 42.500 e 1.955 sequências. O alinhamento PF00576 foi filtrado com parâmetros 0.8 de cobertura e 0.9 de identidade máxima, já para PF00001, pelo fato do número muito alto de sequências foi utilizado parâmetros um pouco mais rigorosos: 0.7 de cobertura e 0.8 de identidade máxima.

Resultados e Discussão

As duas famílias de proteínas utilizadas nas análises foram escolhidas por possuírem características distintas. A PF00576 consiste de uma pequena família composta por basicamente duas subclasses funcionais: a transtiretina, responsável pelo transporte dos hormônios tireoidianos T3 e T4 [Richardson, 2015]; e a Hidrolase 5-hidroxi iso-hidratada (HIUase), enzima presente desde bactérias a vertebrados, envolvida no metabolismo do ácido úrico [Richardson, 2015]. A família dos GPCRs de classe A (Pfam: PF00001) constitui-se da maior classe dos GPCRs, contendo 689 membros em humanos e atualmente 1.827 sequências depositadas no Swiss-Prot, além de centenas de subfamílias classificadas de acordo com relações evolutivas, funcionais e especificidade em relação ao ligante [Joost & Methner, 2002; Southan et al., 2015; Munk et al., 2016]. Além disto, a família possui diversas subfamílias órfãs, proteínas cujo sua função ou ligante ainda são desconhecidos. Muitas dessas sequências orfãs possuem baixa similaridade global de sequência com outras GPCRs já conhecidas, o que as tornam um caso interessante para análise de coevolução [Song et al., 2017].

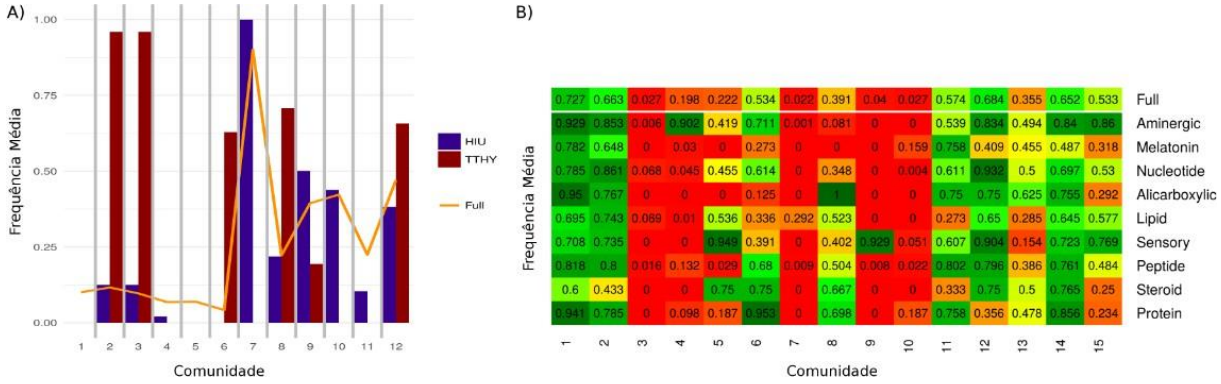

Figura 1: Frequência média dos resíduos de cada comunidade em cada subalinhamento composto por sequências depositadas no Swiss-Prot.

Como é possível observar na figura 1A, o método detectou corretamente comunidades determinantes de especificidade tanto para a classe das HIUases quanto para as Transtiretinas. A comunidade 7, composta pelos nós: Hidroxílas31, Hidrofóbicos32, Hys35, Hidrofóbicos296, Polar327 e Positivamente Carregados328; é completamente conservada entre as HIUases e nula entre as transtiretinas. Já as comunidades 2, 3 e 7, compostas por Val34, Val223, Tyr129, Trp233, Leu30, Met31, Lys35, Phe107, Glu120, His123, His269, Ala296, Gly303, Ala315 e Thr330 são as principais determinantes da classe das transtiretinas.

Ao analisar a distribuição estrutural destes resíduos (Figura 2), é possível observar que em ambos os casos, eles tendem a estarem localizados na região dos sítios de atividade de suas proteínas. Principalmente em regiões que sofreram alterações na estrutura secundária, onde na HIUase se observa presença de alças, e na transtiretinas observa-se a presença de folhas beta, fortalecendo a hipótese de

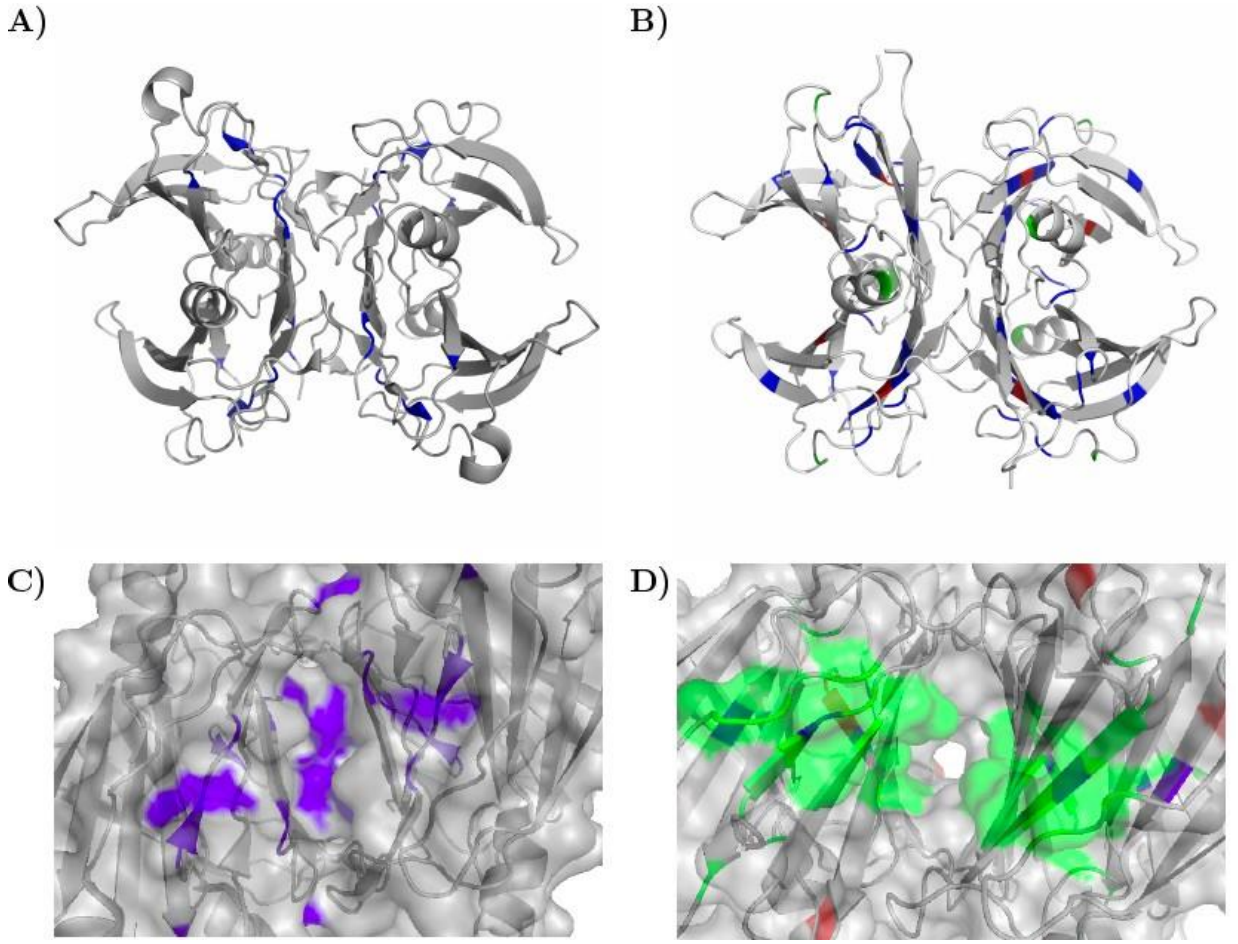

Figura 2 – A e C ilustram a estrutura da HIUase de *Zebrafish* (PDB: 2H1X). B e D mostram a estrutura Transtiretina de rato (PDB: 1GKE). Em ambos os casos as respectivas comunidades determinantes de especificidade estão destacada por uma coloração. Observa-se em C a cavidade do sítio fechado pela Tyr327, detectada na comunidade determinante de HIUase.

que estes resíduos sejam realmente determinantes da divergência funcional entre ambas as classes. Além disto, a presença de um resíduo polar na posição 327 já foi relacionada a importância funcional e evolutiva das HIUases em relação a Transtiretina. Como é possível observar na figura 2C e 2D, a presença deste resíduo fecha completamente o acesso a parte interna do tunel catalítico das HIUases, já na transtiretina este fecho é substituído pelas cadeias secundárias da Thr330 de ambos os dímeros [Cendron et al., 2011] . Ambos os resíduos foram detectados em suas respectivas comunidades de determinantes.

A classificação conseguiu distinguir todas as sequências de HIUases e transtiretinas depositadas no Swiss-Prot, o que leva a crer que as comunidades de resíduos detectadas realmente são suficientes para definir a especificidade funcional destas proteínas.

Na figura 1 também pode-se observar que o método foi eficiente em detectar comunidades que determinam a especificidade de classes gerais das GPCRs, como aminérgicas, alicarboxílicas, sensoriais e proteícas. Além disso, três comunidades poderam ser relacionadas a receptores específicos (Figura 3), como a comunidade 7 aos receptores de prostanóide, comunidade 9 aos receptores de opsínas e comunidade 10 aos receptores de hormônios glicoproteícos.

Outro tipo de padrão possível de observar nas fíguras 1 e 3 é do grupos de resíduos proibitivos para determinadas subfamílias do alinhamento. Este tipo de padrão pode ser tão informativo quanto o de grupos

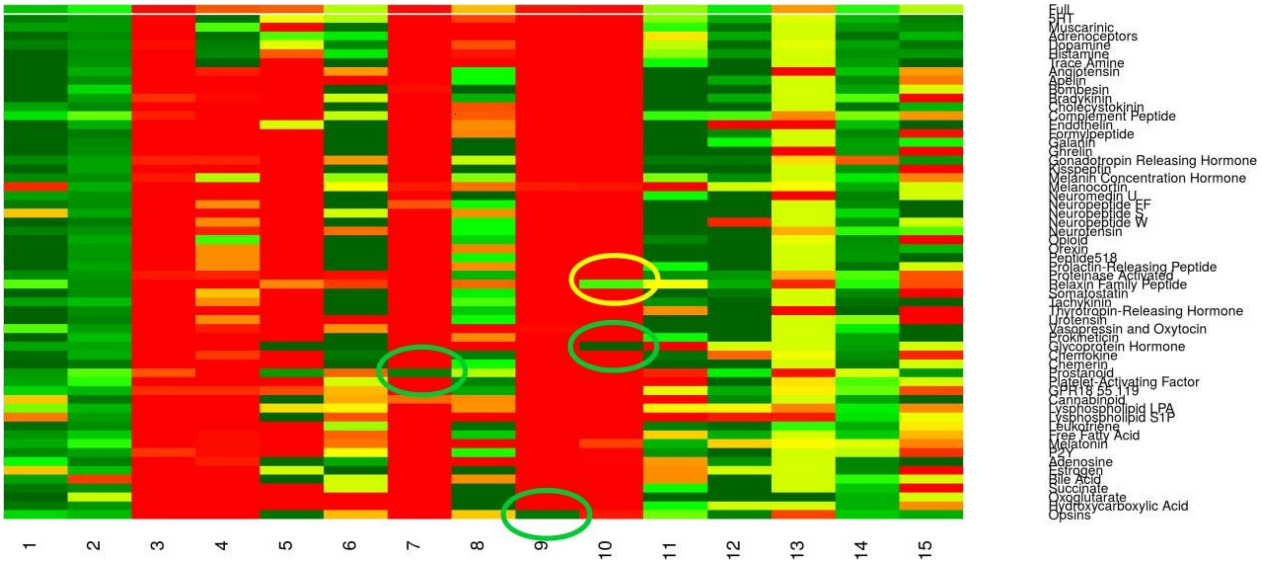

Figura 3 – Frequência média dos resíduos identificados em cada comunidade aos receptores específicos das GPCRs depositados no Swiss-Prot

específicos, sendo assim possível criar uma espécie de assinatura para as sequências de acordo com os vetores de frequência média. Caso estas assinaturas contenham realmente informações acerca de resíduos determinantes e proibitivos das subclasses, seria então possível classificar corretamente até mesmo classes que não obtiveram comunidades relacionadas. Isso foi confirmado durante a classificação, o SVM foi capaz de classificar classes a quais não haviam sido detectadas comunidades determinantes, como das GPCRs baseadas em ligantes esteroides, lipídios e peptídeos, com um F1 score médio de 0.85. Na classificação dos ligantes específicos, apenas 10 das 52 classes avaliadas não mostraram nenhum sinal de especificidade (Aminas Traço, Apelina, Chemerina, Dopamina, Galanina, Lisofosfolipídeo LPA, Neuropeptídeo FF, Neuropeptídeo W, Neurotensina e Urotensina), com um F1 score médio de 0.77.

Foi realizado um teste de classificação com as 195 sequências anotadas como órfãs no GPCRDB [Munk et al., 2016]. É possível observar na figura 4, que a acurácia em relação aos ligantes específicos foi extremamente insatisfatória. Porém tal resultado, além de esperado é interessante, uma vez que estas proteínas possuem ligantes/função desconhecidas, seus rótulos correspondentes não estariam disponíveis na etapa de treinamento. Já em relação as classes genéricas (aminérgicas, esteróides, lípidos, etc.), a classificação obteve resultados instigantes. Mais de 25% das sequências obtiveram probabilidade de estimação acima de 80%.

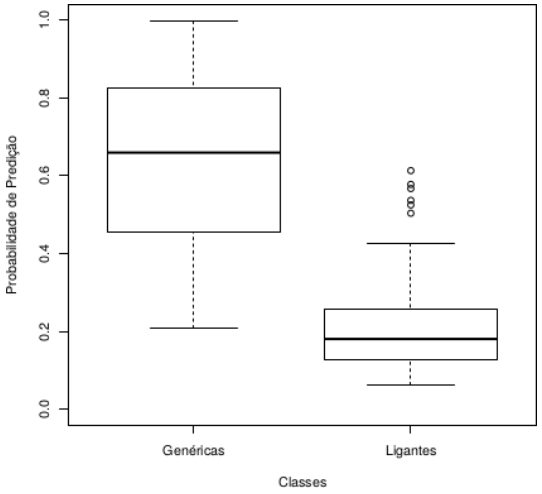

Figura 4 - Distribuição das probabilidades de classificação entre classes genéricas e ligantes específicos das GPCRs órfãs

Todos os receptores órfãos de aminas traço foram corretamente classificado como aminérgicos, além disto foi observado outra relação interessante a respeito da GPR61, pois ainda se sabe muito pouco a seu respeito. Estudos por análises filogenéticas a classificaram como um receptor de melatonina [Bjarnadóttir et al., 2006; Gloriam et al., 2007; Civelli et al., 2013], porém um estudo recente demonstrou sua incapacidade de interação com ligantes de melatonina [Oishi et al., 2017]. Esta sequência possui algumas comunidades

determinantes de especificidade para receptores aminérgicos, e sua probabilidade de predição foi de 0.75. O que nos leva a um palpite que se trate na verdade de um receptor aminérgico.

Conclusão

Os métodos propostos neste artigo contribuem tanto para pesquisadores que trabalham com análises de proteínas, quanto para cientistas de redes em geral, uma vez que essa metodologia pode ser adaptada para resolver problemas em diversos contextos envolvendo redes de coafiliação. A metodologia foi capaz de mapear de forma eficiente e escalável padrões coocorrência entre resíduos e propriedades de proteínas evolutivamente próximas, podendoser aplicada em diversos tipos de pesquisas como identificação de alvos no desenvolvimento de fármacos, estudos evolutivos em geral estudos de caracterização de novas proteínas e predição de efeito de mutações. Além disso, foi possível classificar sequências com uma boa taxa de acerto utilizando os resultados obtidos, o que pode ser futuramente ampliado para uma nova forma de anotação de proteínas.

Bibliografia

Bennett, J. M., & Kendrew, J. C. (1952). The computation of Fourier synthesis with a digital electronic calculating machine. *Acta Crystallographica*, *5*(1), 109–116.

Bjarnadóttir, T. K., Gloriam, D. E., Hellstrand, S. H., Kristiansson, H., Fredriksson, R., & Schiöth, H. B. (2006). Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse. *Genomics*, *88*(3), 263–273. https://doi.org/10.1016/j.ygeno.2006.04.001

Bleicher, L., Lemke, N., Garratt, R. C., Gobel, U., Sander, C., Schneider, R., … Goodwin, D. (2011). Using Amino Acid Correlation and Community Detection Algorithms to Identify Functional Determinants in Protein Families. *PLoS ONE*, *6*(12), e27786. https://doi.org/10.1371/journal.pone.0027786

Cendron, L., Ramazzina, I., Percudani, R., Rasore, C., Zanotti, G., Berni, R. (2011). Probing the evolution of hydroxyisourate hydrolase into transthyretin through active-site redesign. *Journal of molecular biology*, *409*(4), 504-512. https://doi.org/10.1016/j.jmb.2011.04.022

Chakrabarti, S., H.Bryant, S., & R.Panchenko, A. (2007). Functional Specificity Lies within the Properties and Evolutionary Changes of Amino Acids. *Journal of Molecular Biology*, *373*(3), 801–810. https://doi.org/10.1016/J.JMB.2007.08.036

Chakraborty, A., & Chakrabarti, S. (2015). A survey on prediction of specificity-determining sites in proteins. *Briefings in Bioinformatics*, *16*(1), 71–88. https://doi.org/10.1093/bib/bbt092

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the Functional Effect of Amino Acid

Substitutions and Indels. *PLoS ONE*, *7*(10), e46688. https://doi.org/10.1371/journal.pone.0046688

Civelli, O., Reinscheid, R. K., Zhang, Y., Wang, Z., Fredriksson, R., & Schiöth, H. B. (2013). G Protein–Coupled Receptor Deorphanizations. *Annual Review of Pharmacology and Toxicology*, *53*(1), 127–146. https://doi.org/10.1146/annurev-pharmtox-010611-134548

da Fonseca, N. J., Lima Afonso, M. Q., Pedersolli, N. G., de Oliveira, L. C., Andrade, D. S., & Bleicher, L. (2017). *Sequence, structure and function relationships in flaviviruses as assessed by evolutive aspects of its conserved non-structural protein domains*. *Biochemical and Biophysical Research Communications*. https://doi.org/10.1016/j.bbrc.2017.01.041

Dima, R. I., & Thirumalai, D. (2006). Determination of network of residues that regulate allostery in protein families using sequence analysis. *Protein Science*, *15*(2), 258–268. https://doi.org/10.1110/ps.051767306

Fonseca-Júnior, N. J., Afonso, M. Q. L., Oliveira, L. C., & Bleicher, L. (2018). PFstats: A Network-Based Open Tool for Protein Family Analysis. *Journal of Computational Biology*, *25*(5). https://doi.org/10.1089/cmb.2017.0181

Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*. https://doi.org/10.1016/j.physrep.2016.09.002

Gloriam, D. E., Fredriksson, R., & Schiöth, H. B. (2007). The G protein-coupled receptor subset of the rat genome. *BMC Genomics*, *8*. https://doi.org/10.1186/1471-2164-8-338

Joost, P., & Methner, A. (2002). Phylogenetic analysis of 277 human G-protein-coupled receptors as a tool for the prediction of orphan receptor ligands. *Genome Biology*, *3*(11), RESEARCH0063. https://doi.org/10.1186/gb-2002-3-11-research0063

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R., Wyckoff, H., & Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, *181*(4610), 662–666.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, *217*(5129), 624–626. https://doi.org/10.1038/217624a0

Lima Afonso, M., de Lima, L., & Bleicher, L. (2013). Residue correlation networks in nuclear receptors reflect functional specialization and the formation of the nematode-specific P-box. *BMC Genomics*, *14*(Suppl 6), S1. https://doi.org/10.1186/1471-2164-14-S6-S1

Munk, C., Isberg, V., Mordalski, S., Harpsøe, K., Rataj, K., Hauser, A. S., … Gloriam, D. E. (2016). GPCRdb: the G protein-coupled receptor database – an introduction. *British Journal of Pharmacology*. https://doi.org/10.1111/bph.13509

Pazos, F., & Bang, J.-W. (2006). Computational Prediction of Functionally Important Regions in Proteins. *Current Bioinformatics*, *1*(1),

15–23. https://doi.org/10.2174/15748930675330633

Pedruzzi, I., Rivoire, C., Auchincloss, A. H., Coudert, E., Keller, G., De Castro, E., … Bridge, A. (2015). HAMAP in 2015: Updates to the protein family classification and annotation system. *Nucleic Acids Research*, *43*(D1), D1064–D1070. https://doi.org/10.1093/nar/gku1002

Richardson, S. J. (2014). Tweaking the structure to radically change the function: The evolution of transthyretin from 5-hydroxyisourate hydrolase to triiodothyronine distributor to thyroxine distributor. *Frontiers in Endocrinology*, *5*(DEC). https://doi.org/10.3389/fendo.2014.00245

Rios-Anjos, R. M., de Lima Camandona, V., Bleicher, L., & Ferreira-Junior, J. R. (2017). Structural and functional mapping of Rtg2p determinants involved in retrograde signaling and aging of Saccharomyces cerevisiae. *PloS One*, *12*(5).

Schnoes, A. M., Brown, S. D., Dodevski, I., & Babbitt, P. C. (2009). Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology*, *5*(12). https://doi.org/10.1371/journal.pcbi.1000605

Song, J. S., Gonzales, N. R., Yamashita, R. A., Bryant, S. H., & Marchler-Bauer, A. (2017). CDD: functional insights into orphan GPCRs via subfamily domain architectures. *AACR*.

Southan, C., Sharman, J. L., Benson, H. E., Faccenda, E., Pawson, A. J., Alexander, S. P. H., … Davies, J. A. (2016). The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Research*, *44*(D1), D1054–D1068. https://doi.org/10.1093/nar/gkv1037

Suhadolnik, M. L. S., Salgado, A. P. C., Scholte, L. L. S., Bleicher, L., Costa, P. S., Reis, M. P., … Nascimento, A. M. A. (2017). Novel arsenic-transforming bacteria and the diversity of their arsenic-related genes and enzymes arising from arsenic-polluted freshwater sediment. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-11548-8

Tumminello, M., Miccichè, S., Lillo, F., Piilo, J., & Mantegna, R. N. (2011). Statistically validated networks in bipartite complex systems. *PLoS ONE*, *6*(3). https://doi.org/10.1371/journal.pone.0017994

Zuckerkandl, E., & Pauling, L. (1962). Molecular disease, evolution and genetic heterogeneity. *Horizons in Biochemistry*, 189–225.

Zuckerkandl, E., & Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, *8*(2), 357–366. https://doi.org/10.1016/0022-5193(65)90083-4

Financiamento

Este trabalho foi financiado pela

FAPEMIG, CAPES e CNPq.