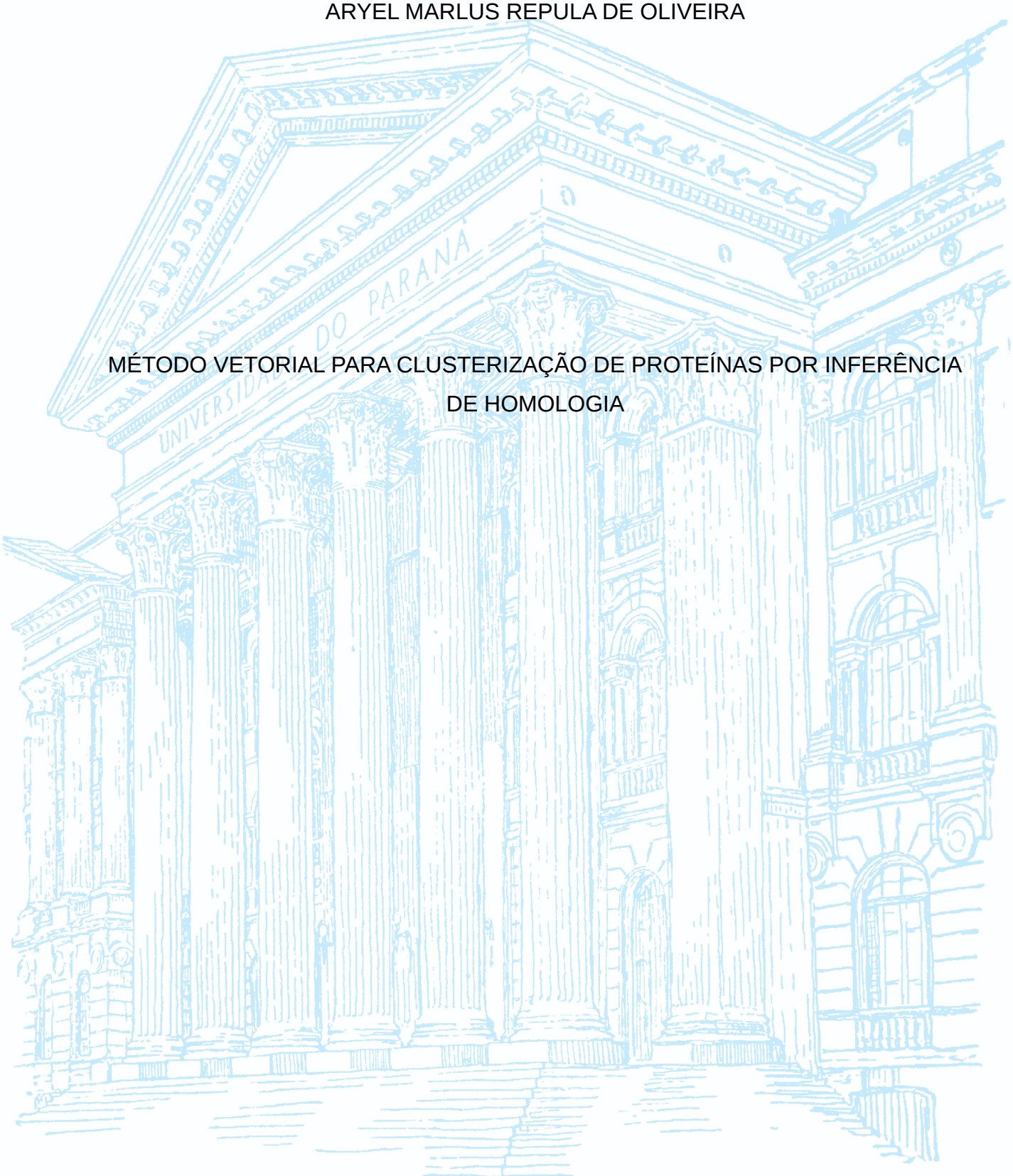
UNIVERSIDADE FEDERAL DO PARANÁ

ARYEL MARLUS REPULA DE OLIVEIRA



MÉTODO VETORIAL PARA CLUSTERIZAÇÃO DE PROTEÍNAS POR INFERÊNCIA DE HOMOLOGIA

CURITIBA

2018

ARYEL MARLUS REPULA DE OLIVEIRA

MÉTODO VETORIAL PARA CLUSTERIZAÇÃO DE PROTEÍNAS POR INFERÊNCIA DE HOMOLOGIA

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Bioinformática, no curso de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica da Universidade Federal do Paraná.

Orientador: Prof. Dr. Roberto Tadeu Raittz

CURITIBA

2018

Catalogação na publicação
Sistema de Bibliotecas UFPR
Biblioteca de Educação Profissional e Tecnológica

O48 Oliveira, Aryel Marlus Repula de
Método vetorial para clusterização de proteínas por inferência de homologia / Aryel Marlus Repula de Oliveira. - Curitiba, 2018.
31 p.: il., tabs, grafs.

Orientador: Roberto Tadeu Raittz
Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Curso de Pós-Graduação em Bioinformática.

1. Homologia (Biologia). 2. Proteínas. 3. Bioinformática. I. Raittz, Roberto Tadeu. II. Título. III. Universidade Federal do Paraná.
CDD 005.369

Elaboração: Angela Pereira de Farias Mengatto - CRB 9/1002

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA

Pós-Graduação em Bioinformática WWW.BIOINFO.UFPR.BR
E-mail: bioinfo@ufpr.br **Tel: 41 33614906**

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em BIOINFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **ARYEL MARLUS REPULA DE OLIVEIRA** intitulada: **Método vetorial para clusterização de proteínas por inferência de homologia**, após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua Aprovação no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 22 de março de 2018.

Dr. Roberto Tadeu Raittz
Presidente
Programa de Pós-graduação em Bioinformática – UFPR

Drª. Leda Satie Chubatsu
Avaliadora Externa
Departamento de Bioquímica - UFPR

Drª. Ana Claudia Bonatto
Avaliador Interno
Departamento de Genética – UFPR

Dr. Dieval Guizelini
Avaliador Interno
Programa de Pós-graduação em Bioinformática - UFPR

À minha filha Ellen, a razão de todo meu esforço.

À minha esposa Suellen, que me incentivou e apoiou para a realização deste trabalho.

AGRADECIMENTOS

À Deus, que, em sua imensa misericórdia, permitiu que fizesse parte deste programa de Mestrado.

À minha filha Ellen, que, quando minhas forças já não eram o suficiente para brincar de pega-pega, teve compreensão e aceitou atividades com menor demanda física.

À minha esposa Suellen, que me incentivou à mudar para a área de pesquisa, mesmo significando começar uma nova carreira.

À todos da família que me ajudaram a percorrer este caminho.

Ao meu orientador Roberto Raittz, que respeito e admiro desde a graduação, foi uma honra trabalhar mais uma vez ao seu lado, é um verdadeiro mestre.

Agradeço a todos os professores do programa de Bioinformática, cada um contribuiu muito para a formação de todos os alunos. Acredito que jamais saberão o impacto real que provocam na vida e na sociedade pois é algo difícil de medir.

À professora Jeroniza e ao professor Dieval, que me ajudaram a tornar este trabalho um pouco mais compreensível, seus esforços em tentar traduzir o que eu tentava dizer nunca serão esquecidos, agradeço pela paciência, força de vontade e disponibilidade.

Agradeço à minha nova orientadora Ana Claudia Bonatto por me ajudar a reestruturar o texto em algumas partes desta dissertação após a defesa. Sou grato também por me acolher em seu laboratório, sei que tenho muito a aprender e espero conseguir colaborar em suas pesquisas.

À todos os alunos que conviveram comigo no laboratório de bioinformática, aprendi muito com todos e espero profundamente que eu também tenha contribuído positivamente com vocês.

”Talvez seja bom ter uma mente bonita, mas um dom ainda maior é descobrir um coração bonito”

John Nash

RESUMO

A identificação de proteínas que possuem algum grau de homologia é fundamental para diversas pesquisas biomédicas, como por exemplo inferência de funções, estrutura, anotação de sequências, e construção de árvores filogenéticas. A clusterização é uma das principais técnicas utilizadas nesta área e os melhores resultados são obtidos utilizando matrizes de similaridade, geradas a partir de alinhamentos entre sequências e modelos probabilísticos de Markov, como entrada para os métodos de clusterização. Porém inevitavelmente estas matrizes propagam suas próprias limitações através dos métodos utilizados posteriormente pois também são heurísticas. Esta dissertação pretende contribuir com a identificação de proteínas homólogas utilizando técnicas de clusterização, criando um método livre de alinhamento para gerar matrizes de similaridade entre sequências com objetivo de minimizar tendências inerentes ao alinhamento. O estudo de caso demonstrou que nosso método possui um resultado mais estável que os representantes do estado da arte considerando as métricas F1-Score, precisão e sensibilidade. Apresentamos também uma estratégia para avaliação automática de clusters de proteínas que superou as limitações do estado da arte, viabilizando a determinação não-supervisionada dos melhores parâmetros para clusterização de sequências de aminoácidos por inferência de homologia.

Palavras-chave: Clusterização, Inferência de Homologia, Proteínas, Bioinformática.

ABSTRACT

Protein homology detection is fundamental for many biomedical researchs, for example, function and structure prediction, sequence annotation, and phylogenetic analysis. Clustering is one of the main techniques used to handle this problem, and the best results make use of similarity matrix, generated by sequence alignment and probabilistic Markov models, as input to clustering algorithms. However, this similarity matrix have their own heuristics limitations that are propagated to the clustering techniques. This work aims to contribute to protein clustering by homology inference, developing a consistent alignment-free similarity matrix, to minimize the alignment limitations. The case study demonstrate that our method make more stable result considering the F1-Score, precision and recall metrics, than the state of art techniques. We develop also a strategy to evaluate and select the most promising clustering result without human supervision, that superate the current limitations of the state of art in this task for biomedical data.

Keywords: Clustering, Homology Detection, Protein, Bioinformatics.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

APRESENTAÇÃO

Esta dissertação está estruturada em formato de artigo e possui três partes principais. A Parte I apresenta a fundamentação teórica necessária para um bom entendimento da área e os objetivos desta dissertação.

A parte II possui duas seções, a primeira contém o rascunho do atigo científico a ser submetido em um periódico da área e apresenta os principais resultados criados por este trabalho. A segunda seção aborda uma discussão mais ampla e apresenta resultados complementares que não fizeram parte do escopo do artigo, mas que são igualmente importantes para compreensão dos métodos criados nesta dissertação.

A parte III apresenta as principais conclusões desta dissertação.

Parte I

FUNDAMENTAÇÃO TEÓRICA

1 FUNDAMENTAÇÃO TEÓRICA

1.1 FINALIDADE DA BUSCA POR PROTEÍNAS HOMÓLOGAS

Proteínas que acredita-se descender de um ancestral comum são consideradas homólogas. A semelhança estrutural e/ou funcional tende a ser mais conservada que a similaridade observada entre sequências de aminoácidos. Identificá-las viabiliza análises filogenéticas, anotação de sequências, inferência de função, estrutura e análises de interação proteína-proteínas (BERNARDES J. S.; VIEIRA; ZAVERUCHA, 2015) (VU D.; SZöKE; ROBERT, 2014).

Inferir homologia utilizando os métodos de alinhamento entre sequências atuais só é possível (sem ambiguidades) quando a identidade observada é maior que 40% em um longo trecho de aminoácido. Pares aquém deste percentual que compartilham ancestralidade são descobertas por métodos que possuem maior custo computacional ou experimental e envolvem uso de informações estruturais. Estas limitações são um desafio para inferência de homologia pois os dados estruturais crescem na ordem de milhares ao ano enquanto a inclusão de sequências de aminoácidos ocorre na ordem de milhões ao mês (CHEN et al., 2018).

1.2 PRINCIPAIS BANCOS DE DADOS DE PROTEÍNAS, E SEUS MÉTODOS PARA INFERÊNCIA DE HOMOLOGIA

O agrupamento de proteínas (clusterização) é um dos principais métodos para realizar inferência de homologia (ROTTGER R.; KALAGHATGI; BAUMBACH, 2013). Os bancos de dados biológicos utilizam diferentes métodos e critérios para gerar seus clusters e muitos deles são integrados para viabilizar a comparação entre seus resultados. Os principais bancos de dados e seus métodos são apresentados a seguir.

1.2.1 UNIPROT KNOWLEDGEBASE (UNIPROTKB)

O UniProtKB possui anotações e informações funcionais de proteínas. O método utilizado é procurar domínios conhecidos em sequências desconhecidas (CONSORTIUM, 2016). É dividido em duas seções:

- Revisada (Swiss-Prot): O resultado do processo computacional de inferência é validado por especialistas e utiliza informações da literatura para fundamentar as predições. Possui 556.825 proteínas (em 03/2018).
- Não-revisada (TrEMBL): Possui 108.857.716 proteínas (em 03/2018) anotadas automaticamente utilizando uma das duas estratégias:
  - Verificação da presença de domínios funcionais conhecidos utilizando informações obtidas de outros bancos de dados (InterPro, Pfam, PROSITE e SMART);
  - Caso a etapa anterior não apresente resultados, utilização de um modelo probabilístico e de aprendizagem de máquina para encontrar os domínios funcionais estatisticamente mais próximos aos trechos presentes nas sequências (geralmente são anotações com menor confiabilidade que a primeira abordagem).

1.2.2 UNIPROT REFERENCE CLUSTERS (UNIREF)

O banco de dados UniRef disponibiliza sequências de aminoácidos agrupadas com grau de identidade de alinhamento global em 100%, 90% e 50%. Ele contém as proteínas presentes no UniProtKB e algumas selecionadas do UniProt Archive (que possui a maioria das sequências disponíveis publicamente no mundo) (SUZEK et al., 2015).

Os grupos de sequências de aminoácidos são criados utilizando o software CD-HIT (SUZEK et al., 2015). O método dele calcula a similaridade de alinhamento global do conjunto em relação às sequências definidas como referência (seed) para cada cluster conforme a Figura 1 (LI; GODZIK, 2006). No caso do UniRef é definido o grau de identidade (100%, 90% ou 50%) com uma sobreposição (overlap) mínima de 80% em relação à *seed* (SUZEK et al., 2015).

1.2.3 PFAM, SIMPLE MODULAR ARCHITECTURE RESEARCH TOOL (SMART), PROSITE E THE INSTITUTE FOR GENOMIC RESEARCH'S DATABASE OF PROTEIN FAMILIES (TIGRFAM)

Os bancos de dados Pfam (FINN et al., 2015), Simple Modular Architecture Research Tool (SMART) (LETUNIC; BORK, 2017), PROSITE (SIGRIST et al., 2002) e The Institute

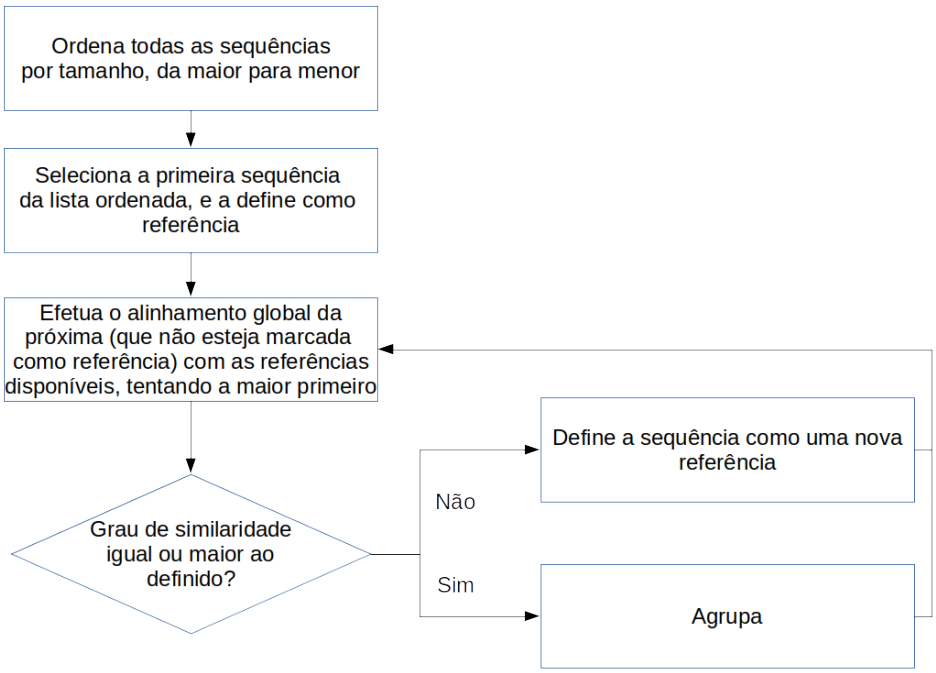


Figura 1: Simplificação da estratégia utilizada pelo algoritmo do software CD-HIT.

for Genomic Research's database of protein families (TIGRFAM) (HAFT et al., 2012) utilizam inferência de domínios conservados para definir seus clusters.

Eles foram agrupados neste tópico por adotarem essencialmente a mesma estratégia (Figura 2) e compartilharem algumas funcionalidades de busca. A principal diferença entre eles está no protocolo utilizado para determinar as proteínas que serão alinhadas. A ferramenta HMMER (MISTRY J.; FINN; PUNTA, 2013) é uma referência para busca de homólogos (SARIPELLA G. V.; SONNHAMMER; FORSLUND, 2016) utilizando alinhamento múltiplo e Modelos Ocultos de Markov (Hidden Markov Models, HMM) e é utilizada por estes bancos de dados. O TIGRFAM utiliza também o PSI-BLAST como alternativa ao HMMER.

1.2.4 NCBI PROTEIN CLUSTERS (PRK), ORTHODB, ORTHOMCLDB E CLUSTERS OF ORTHOLOGOUS GROUPS (COG/KOG)

Estes bancos de dados disponibilizam grupos de proteínas criados por inferência de ortologia (um dos tipos homologia). O método utilizado consiste em efetuar alinhamento local entre sequências de aminoácidos e utilizar o resultado para criar os clusters. As diferenças entre eles se deve ao uso de ferramentas:

- NCBI Protein Clusters (PRK): Os grupos são gerados a partir do melhor hit recíproco obtido da ferramenta NCBI BLAST (CAMACHO C.; COULOURIS; MADDEN, 2009).

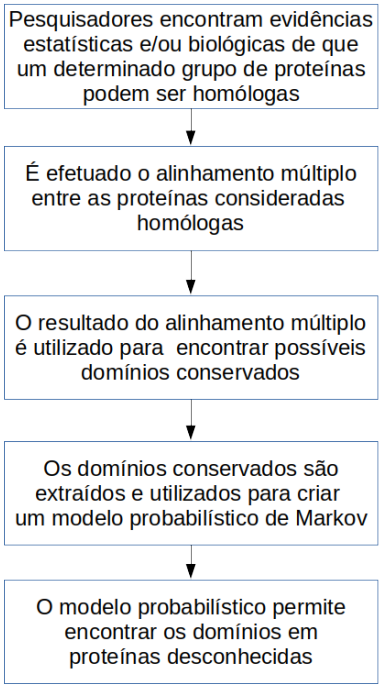


Figura 2: Estratégia utilizada pelos bancos de dados Pfam, SMART e PROSITE.

Posterior a este processo é feita uma clusterização hierárquica que permite verificar clusters relacionados ou próximos. O software USEARCH (EDGAR, 2010) é utilizado para obter o centro (*centroid*) de cada grupo para realizar este último processo.

- OrthoMCLDB (CHEN et al., 2006) , OrthoDB (KRIVENTSEVA et al., 2014) e COG/KOG (KRISTENSEN et al., 2010) : utilizam os valores gerados pelo NCBI BLAST (exceto o OrthoDB que utiliza o equivalente SWIPE) como entrada para um algoritmo de agrupamento. Respectivamente o Markov Clustering (MCL) (ENRIGHT A. J.; DONGENAND; OUZOUNIS, 2002), CD-HIT e COGsoft (mais rápido que o MCL, segundo seus criadores). A validação dos resultados é feita manualmente e ao final do processo é realizado um alinhamento múltiplo para obter domínios conservados para cada um dos clusters.

1.2.5 NCBI CONSERVED DOMAINS DATABASE (CDD)

É um repositório de domínios conservados presentes nos bancos de dados Pfam, SMART, COG/KOG, TIGRFAM, PRK e NCBI-Curated Domains. Nos domínios curados (NCBI-Curated Domains) utiliza informações sobre inferência de estrutura 3D para melhorar a inferência de homologia e função (MARCHLER-BAUER et al., 2016).

1.2.6 RESUMO DOS MÉTODOS MAIS UTILIZADOS

O principal método utilizado pelos bancos de dados para classificar sequências de aminoácidos é a inferência de domínios conservados. Esta inferência é feita a partir de alinhamento múltiplo entre proteínas homólogas agrupadas conforme ilustrado na Figura 3.

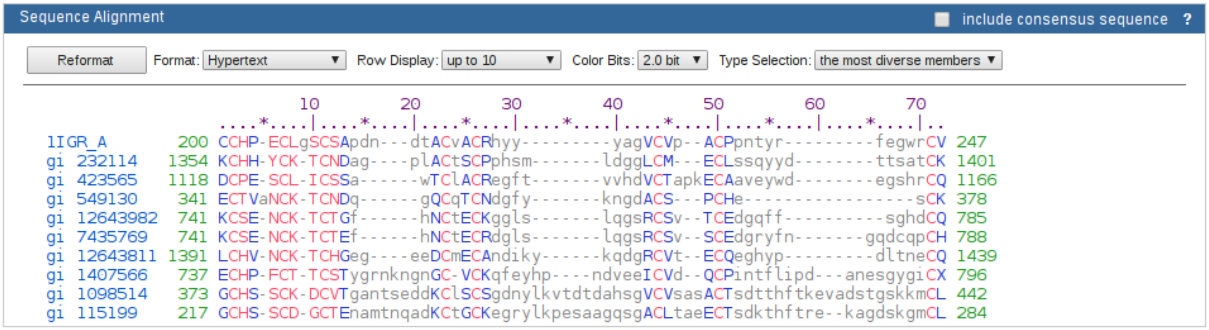


Figura 3: Alinhamento múltiplo parcial obtido do banco de dados CDD efetuando a busca pela família *Furin-like repeats* (ID 238021, em 05/2018). Cada linha é a sequência encontrada em diferentes organismos. Em maiúsculo são apresentados os resíduos conservados em uma escala de azul (menos conservado) e vermelho (mais conservado). Em minúsculo e cinza estão representados os resíduos não-alinhados. O traços em cinza apresentam os *gaps* gerados pelo alinhamento.

O agrupamento (clusterização) de sequências de aminoácidos homólogas é portanto a base para os principais bancos de dados. Erros no agrupamento propagam inconsistências ao processo e são corrigidos em novas versões disponibilizadas pelos bancos. As limitações desta atividade podem ser percebidas observando as estatísticas de domínios conservados apresentadas na Tabela 1 em relação ao tamanho da base de dados TrEMBL, os domínios conservados conhecidos totalizam menos de vinte mil enquanto as sequências disponíveis ultrapassam cem milhões.

Banco	Número de Domínios
CDD	12805
Pfam	16712
PROSITE	1309
SMART	1312
TIGRFAM	4488

Tabela 1: Número de domínios conservados conhecidos. Os bancos não apresentados não disponibilizam estas estatísticas.

1.3 MÉTODOS PARA CLUSTERIZAÇÃO

A clusterização (agrupamento) é uma área também conhecida como aprendizagem não-supervisionada de máquina ou análise exploratória de dados. Seu objetivo é encontrar

grupos que sejam compostos por objetos semelhantes entre si em um conjunto de dados com classificações desconhecidas (XU; WUNSCH, 2010).

Os métodos para clusterização são classificados quanto à sua estratégia (ANDREOPOULOS B.; SCHROEDER, 2009):

- Particionados: separa os objetos do conjunto em um número definido arbitrariamente pelo pesquisador. Exemplo: K-Means, K-Medoids, Fuzzy C-Means, Fanny.
- Densidade: agrupa os objetos de acordo com a densidade observada. Exemplo: DBScan.
- Grafos: Trata os dados de entrada como nós de grafos e busca encontrar caminhos que os conectem. Exemplos: MCL, COGsoft, TransClust.
- Modelos: assume que os dados observados são uma amostra extraída de modelos estatísticos que os geraram, e tenta inferir estes modelos. Exemplo: Kohonen.
- Hierárquico: considerando as distâncias relativas entre os objetos, cria uma estrutura de agrupamentos organizados hierarquicamente. Exemplo: Linkage Average, Linkage Complete, Spectral Clustering.

Métodos da mesma classe podem apresentar resultados diferentes. A Figura 4 exemplifica resultados distintos entre os métodos K-Means e K-Medoids (mesma classe) em um conjunto de dados criado artificialmente utilizando o mesmo valor de parâmetro (número de clusters $K = 3$) (PARK; JUN, 2009).

Todos os métodos possuem ao menos um parâmetro que determinará como os grupos serão gerados. Os valores ideais para estes parâmetros dependem do conjunto de dados e do objetivo da análise. Eles são frequentemente estimados de forma subjetiva pelos pesquisadores (WIWIE C.; BAUMBACH; RöTTGER, 2015).

1.4 VALIDAÇÃO DE CLUSTERS

A validação de agrupamentos gerados a partir de um método de clusterização é tão importante quanto o próprio método utilizado. Ela apoia a definição do melhor valor para os parâmetros dos algoritmos de forma não-subjetiva, viabilizando aplicações totalmente não-supervisionadas (HASSANI; SEIDL, 2016).

O processo de validação de clusters consiste em calcular índices aos agrupamentos e são divididos em duas categorias (LIU et al., 2010):

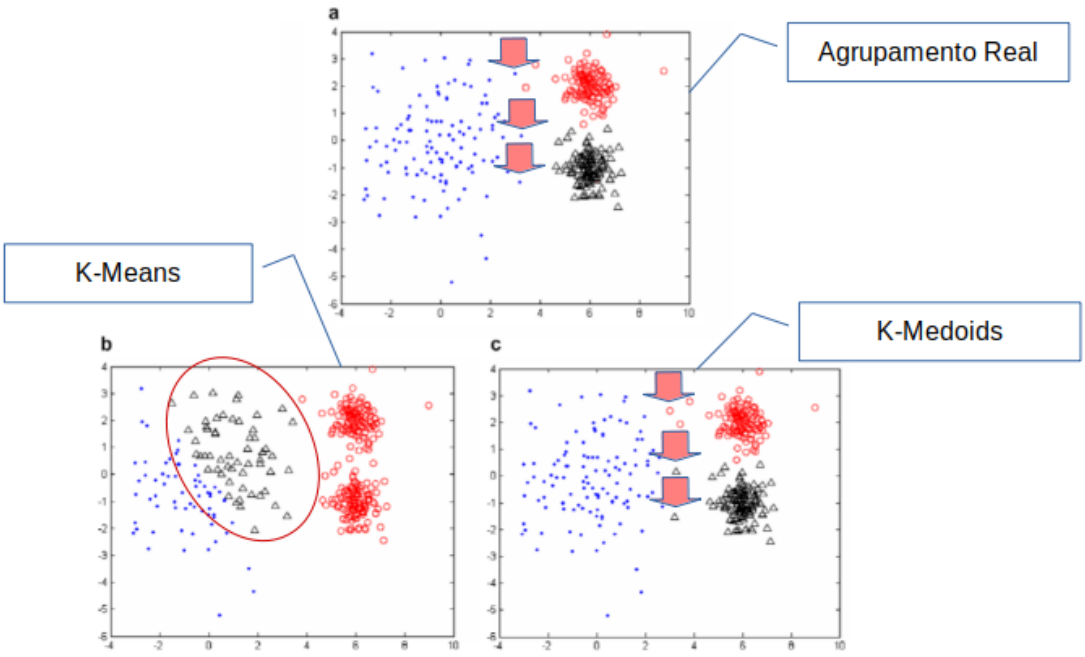


Figura 4: Mesmo utilizando uma estratégia semelhante e os mesmos parâmetros de configuração os métodos K-Means e K-Medoids apresentam resultados diferentes (círculo e setas vermelhas, adaptado de Park et al, 2009).

- Índices externos: Comparam dois resultados de agrupamentos e verificam a concordância entre os clusters.
- Índices internos: Consideram apenas informações do próprio conjunto de dados em seus cálculos.

Os índices internos são a única forma de validação quando os objetos de estudo não possuem nenhuma classificação (LIU et al., 2010). Eles são afetados pelo tipo do conjunto de dados e podem ser contraditórios em relação aos índices externos (WIWIE C.; BAUMBACH; RöTTGER, 2015).

1.4.1 ÍNDICES EXTERNOS

Os índices externos são calculados a partir da comparação entre dois resultados de agrupamentos. Frequentemente um deles é considerado correto (controle, ou padrão-ouro) por especialistas humanos e o outro é resultado de um algoritmo de clusterização a ser avaliado (LIU et al., 2010). São utilizados para estimar performance futura de um método quando aplicado em conjuntos de dados semelhantes ao controle ou para comparar ferramentas (WIWIE C.; BAUMBACH; RöTTGER, 2015).

Os principais índices são derivados dos valores de verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN). Estas variáveis formam a Matriz de Confusão observada na Figura 5. Considerando dois resultados de clusterização C e C', sendo C' considerado o conjunto correto (padrão ouro), a obtenção dos valores para estas variáveis pode ser realizada de duas formas (WIWIE C.; BAUMBACH; RöTTGER, 2015):

- Par a par: todos os membros dos dois agrupamentos (C e C') são comparados entre si. Dado um par de elementos a e b, eles são contabilizados em uma das variáveis da matriz de confusão conforme a definição:
  - Verdadeiro positivo (TP): se os dois elementos estão no mesmo cluster em C e C'.
  - Verdadeiro negativo (TN): se os dois elementos estão em clusters diferentes em C e C'.
  - Falso positivo (FP): se os dois elementos estão no mesmo cluster em C mas em clusters diferentes em C'.
  - Falso negativo (FN): se os dois elementos estão em clusters diferentes em C mas no mesmo cluster em C'.
- Mapeamento: é feito um mapeamento entre os clusters presentes em C e C'. Para cada cluster $c_i \in C$ é selecionado o cluster $c'_j \in C'$ que possua o maior número de elementos em comum: $c_i \cap c'_j \rightarrow max$. Um elemento a será contabilizado em uma das variáveis da matriz de confusão conforme a definição:
  - Verdadeiro positivo (TP): se $a \in c_i \wedge a \in c'_j$.
  - Verdadeiro negativo (TN): 0, pois não há definição nesta abordagem.
  - Falso positivo (FP): se $a \notin c_i \wedge a \in c'_j$.
  - Falso negativo (FN): se $a \in c_i \wedge a \notin c'_j$.

As derivações destas variáveis (TP, TN, FP, FN) geram índices que valorizam diferentes critérios (Figura 5). O $F_1 score$ é a métrica mais aceita para avaliar métodos de clusterização porque utiliza três dessas quatro variáveis em seu cálculo (é a média harmônica entre precisão e sensibilidade), mas é recomendável apresentar sua composição para que o pesquisador decida qual critério é mais relevante para sua análise (LEVER J.; KRZYWINSKI; ALTMAN, 2016). A Figura 6 exemplifica como as métricas podem apresentar valores contraditórios entre si e ocultar comportamentos importantes ao pesquisador.

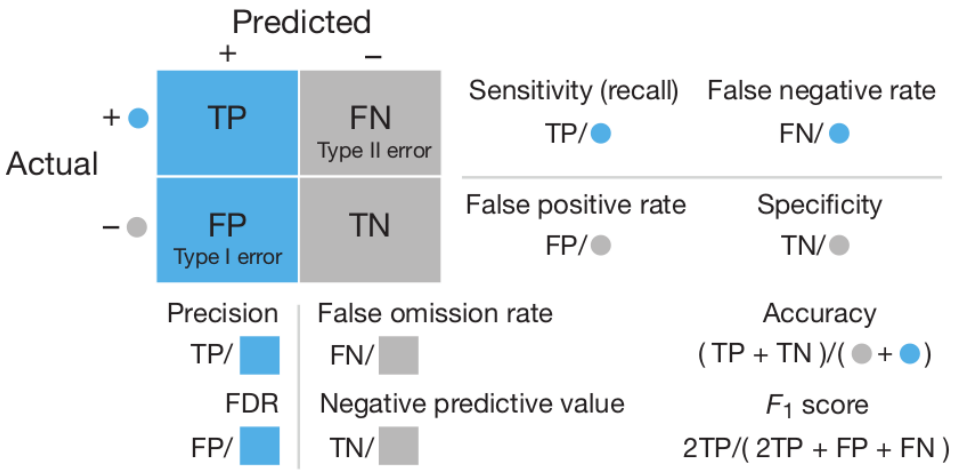


Figura 5: Matriz de confusão e índices que derivam dela. (TP) verdadeiro positivo. (FP) falso positivo. (TN) verdadeiro negativo. (FN) falso negativo (extraído de Lever et al, 2016).

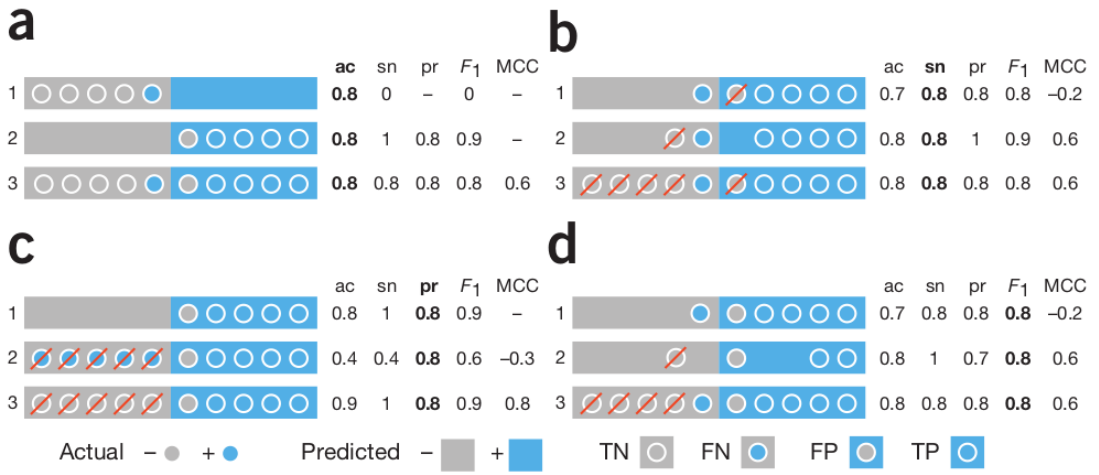


Figura 6: O valor em negrito apresenta a métrica avaliada (a-d) sempre em 0.8 e demonstra as diferentes configurações do resultado (1-3). Em cada painel a observação que não contribui para o cálculo da métrica em negrito está com uma linha vermelha. (ac) acurácia, (sn) sensibilidade, (pr) precisão, (F1) F1-Score, (MCC) Matthews correlation coefficient (adaptado de Lever et al, 2016).

1.4.2 ÍNDICES INTERNOS

Aplicações práticas de clusterização utilizam objetos com classificações desconhecidas. As informações disponíveis para cálculos nestes casos são as presentes no próprio conjunto de dados. Os índices internos utilizam estas informações para avaliar resultados e portanto são a única opção para validação de clusters em aplicações realmente não-supervisionadas (LIU et al., 2010).

Os principais índices internos avaliam a estrutura dos clusters gerados em relação à compactação e separação (HASSANI; SEIDL, 2016):

- Compactação: Avalia o quão próximos os objetos de um cluster estão estre si (Figura 7).

- Separação: Verifica quão separados os clusters estão entre si (Figura 8).

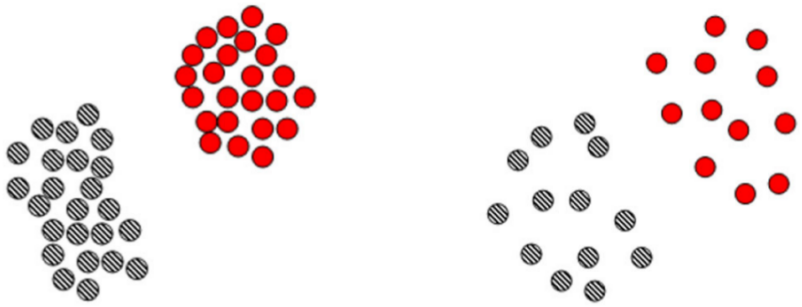


Figura 7: Compactação: Os clusters da esquerda estão mais compactados que os da direita e são considerados melhores pelos índices (adaptado de Hassani et al, 2016).

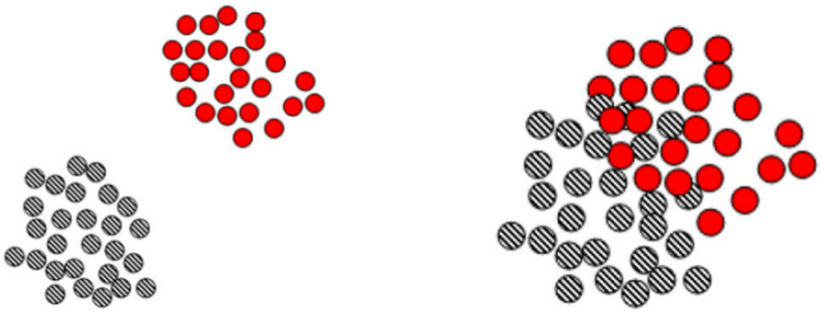


Figura 8: Separação: Os clusters da esquerda estão mais separados que os da direita e são considerados melhores pelos índices (adaptado de Hassani et al, 2016).

Os índices internos diferem entre si em relação a como são calculados os valores para compactação e separação. Os principais utilizados em dados biomédicos são definidos simplificadamente como (WIWIE C.; BAUMBACH; RöTTGER, 2015):

- Davies Bouldin: Considera a distância média entre os membros de cada grupo em relação ao seu núcleo (centroid).
- Dunn: Verifica o diâmetro máximo de um grupo em relação à distância mínima entre os outros grupos.
- Silhouette: Para cada objeto do conjunto é medido o quão similar ele é em relação à todos os objetos do seu grupo, quando comparado com os objetos no grupo mais próximo.
- Calinski Harabasz: Verifica se a variância observada entre os membros dentro de um grupo é pequena e se a variância observada entre os membros de dois grupos diferentes é grande.

O tipo de distribuição do conjunto de dados afeta a eficácia dos índices internos e muitas vezes geram valores contraditórios aos índices externos (WIWIE C.; BAUMBACH; RöTTGER, 2015). O formato dos dados pode ser influenciado ao se definir atributos que não representem bem os objetos reais, pelo uso de uma amostra muito pequena da realidade (baixa dimensionalidade) ou pela inclusão de dados não relacionados ao objetivo (inserção de ruído). A Figura 9 exemplifica um cenário onde a conformação dos dois clusters (vermelho e preto) não favorece os critérios de compactação e separação.

Selecionar o melhor resultado de clusterização utilizando apenas índices internos em dados conhecidos de expressão gênica, similaridade entre sequências de aminoácidos e similaridade estrutural entre proteínas é atualmente inviável. A distribuição do conjunto de dados biológicos é complexa e os índices atuais não são suficientes para determinar *a priori* as classificações de forma totalmente não-supervisionada (WIWIE C.; BAUMBACH; RöTTGER, 2015).

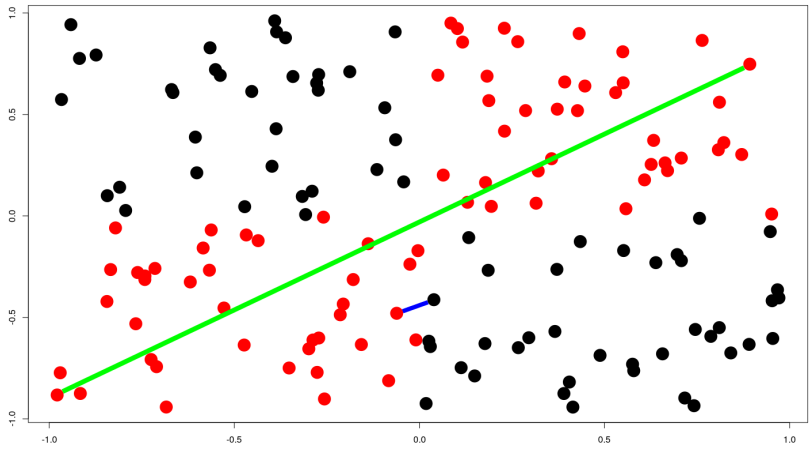


Figura 9: A distribuição destes dados não favorece os critérios de compactação e separação. A distância entre objetos do mesmo cluster (linha verde) é maior que a distância entre objetos de clusters diferentes (linha azul) em vários casos. Considerando que esta é a conformação ideal dos clusters, os índices internos serão contraditórios em relação aos externos (o autor).

1.5 VALIDAÇÃO DE NOVOS MÉTODOS DE CLUSTERIZAÇÃO

Validar novos métodos para clusterização exige sua execução em uma base de dados com classificações consolidadas, e o cálculo de índices externos em relação à ela (WIWIE C.; BAUMBACH; RöTTGER, 2015).

1.5.1 BASE DE DADOS DE REFERÊNCIA PARA PROTEÍNAS HOMÓLOGAS

A classificação de proteínas homólogas é frequentemente resultado de um processo de inferência e portanto está sujeita a erros (SARIPELLA G. V.; SONNHAMMER; FORSLUND, 2016). Estes erros são evidenciados observando as correções realizadas pelos bancos de dados a cada nova versão. Portanto a escolha de uma referência confiável para avaliar novos métodos para clusterização de sequências de aminoácidos é um desafio.

Para minimizar este problema foi criada uma base de dados para validação de homologia entre proteínas que possui duas seções (BROWN et al., 2006)

- *GOLD*: Possui sequências de aminoácidos de enzimas que tiveram sua atividade biológica validada experimentalmente. O grau de similaridade entre alinhamento de proteínas da mesma família é frequentemente baixo (menores que 50%).
- *SILVER*: Possui sequências que possuem alta probabilidade de serem homólogas às da seção *GOLD*, porém não foram validadas experimentalmente e portanto apresentam classificações menos confiáveis.

A seção *GOLD* é a principal referência utilizada em comparações entre métodos para clusterização por ser mais confiável (BERNARDES J. S.; VIEIRA; ZAVERUCHA, 2015). Ela possui 866 sequências de aminoácidos atribuídas manualmente à 91 famílias (BROWN et al., 2006).

1.5.2 BASE DE DADOS DE REFERÊNCIA PARA TESTES

A base de dados *Íris* é uma referência clássica para avaliação de clusterizadores, classificadores e métodos estatísticos. Ela possui dados de 150 amostras de flores de três espécies de *Íris* (*Setosa, Virginica e Versicolor*). Cada amostra possui quatro tipos de informações: altura e largura da pétala, altura e largura da sépala em centímetros (FISHER et al., 1936).

1.6 JUSTIFICATIVA

O agrupamento de sequências de proteínas por inferência de homologia é necessário para viabilizar estudos em diversas áreas da genética, bioquímica e biomedicina. Porém existem várias lacunas que precisam ser melhoradas, entre elas:

- Alinhamento: todas as principais estratégias e bancos de dados utilizam alguma forma de alinhamento entre sequências (local, global, múltiplo ou semi-global), porém estas técnicas possuem limitações conhecidas em relação à máximos locais e são diretamente afetadas pelo tamanho do conjunto estudado (DURBIN et al., 2002) (SARIPELLA G. V.; SONNHAMMER; FORSLUND, 2016).

- Avaliação de agrupamentos: Os parâmetros utilizados pelos algoritmos de clusterização são definidos de forma subjetiva, pois o cálculo de métricas internas para determinar *a priori* a melhor configuração para dados biomédicos não é eficaz (WIWIE C.; BAUMBACH; RöTTGER, 2015).

1.7 OBJETIVO GERAL

Este trabalho tem objetivo de contribuir com o avanço dos métodos de agrupamentos de proteínas por inferência de homologia.

1.8 OBJETIVOS ESPECÍFICOS

- Desenvolver um método para agrupamento de proteínas que apresente um F1-Score elevado enquanto mantém a estabilidade entre precisão e sensibilidade, sem utilizar alinhamentos entre sequências.

- Desenvolver um método para avaliação dos agrupamentos de sequências de aminoácidos que viabilize a análise de grupos homólogos *a priori*.

Parte II

Material Produzido

2 MANUSCRITO

2.1 MANUSCRITO AUTORAL A SER SUBMETIDO À UM PERIÓDICO

O rascunho do artigo que será submetido à um periódico da área é apresentado neste capítulo.

2.2 PROTOCOLO DE AVALIAÇÃO: MÉTODOS DE AGRUPAMENTO

Avaliar e comparar métodos de agrupamento envolve a elaboração de estudos de caso que representem problemas relevantes para a área de aplicação proposta. Este capítulo apresenta os critérios e protocolos utilizados para comparar e avaliar os principais métodos de agrupamento.

2.2.1 SELEÇÃO DOS MÉTODOS DE AGRUPAMENTO PARA COMPARAÇÕES DE DESEMPENHO

A seleção dos métodos de agrupamento que foram comparados (Tabela 2) utilizou dois critérios para escolha:

- Foram selecionados métodos que apresentaram métricas externas superiores em relação aos que foram avaliados em benchmarkings publicados sobre clusterização de sequências de aminoácidos.
- Métodos utilizados nos principais bancos de dados de sequências de aminoácidos foram escolhidos para comparação devido sua ampla utilização, embora nem todos podem ser considerados preditores de homologia por definição.

Método	**Utilizado em**
MCL	OrthoMCLDB, e benchmarkings de homologia
TransClust	Re-anotação de fungos, e benchmarkings de homologia
PHMMER	Pfam, CDD, e benchmarkings de homologia
NCBI BLAST	Quase todas as bases de dados, e benchmarkings de homologia
CD-HIT	UniRef, OrthoDB
USEARCH	NCBI Protein Clusters

Tabela 2: Métodos comparados.

Cada método da Tabela 2 possui parâmetros que são definidos pelo pesquisador. Com o objetivo de verificar o potencial de acerto dos métodos em uma base de proteínas homólogas, efetuamos clusterizações variando os parâmetros conforme a Tabela 3 e registramos o maior F1-Score que cada método obteve. O NCBI BLAST e PHMMER foram executados em suas configurações padrão e seus grupos foram gerados pelo método de agrupamento utilizando o melhor hit recíproco (utilizado no banco de dados NCBI Protein Clusters).

Método	Parâmetro	Variação Testada	Step size
MCL	Inflation $I \in [1.1, 20]$	$1.4 : 20$	2
TransClust	Threshold $T \in [1, 100]$	$1 : 100$	5
CD-HIT	Similaridade $S \in [0.4, 1]$	$0.5 : 0.9$	0.1
USEARCH	Similaridade $S \in [0.4, 1]$	$0.5 : 0.9$	0.1

Tabela 3: Variação de parâmetros. n **se refere ao número total de objetos no conjunto**

2.2.2 ESTUDO DE CASO: BASE DE DADOS DE PROTEÍNAS HOMÓLOGAS

A base de dados *GOLD* (BROWN et al., 2006) é padrão-ouro para comparações de inferência de homologia (BERNARDES J. S.; VIEIRA; ZAVERUCHA, 2015)(ROTTGER R.; KALAGHATGI; BAUMBACH, 2013) e suas classificações em nível de família de proteínas foram consideradas como referência para o cálculo das métricas externas F1-Score, precisão e sensibilidade.

2.2.3 FLUXOGRAMA DO PROTOCOLO DE TESTES: MÉTODOS DE AGRUPAMENTO

A Figura 10 possui o fluxograma do protocolo utilizado para efetuar os testes comparativos entre os métodos de agrupamento selecionados (WIWIE C.; BAUMBACH; RöTTGER, 2015).

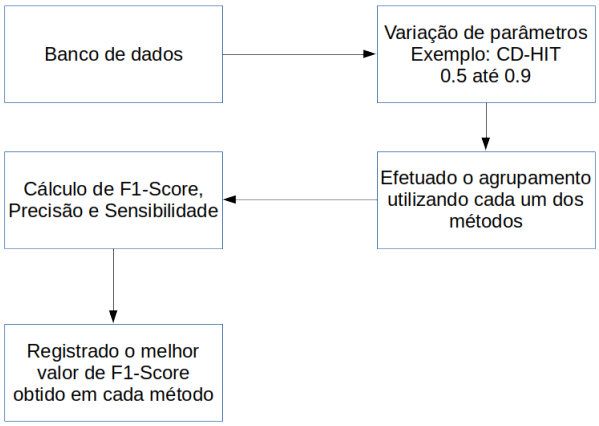


Figura 10: Fluxograma do protocolo de testes comparativos para os métodos de agrupamento

2.3 RESULTADOS COMPLEMENTARES: MÉTODO VETORIAL

A Tabela 4 possui os resultados obtidos na execução do protocolo de testes descrito nas seções anteriores. O método Vetorial para clusterização desenvolvido neste trabalho está definido na seção 2.1 desta dissertação e o protocolo de testes aplicado foi o mesmo para todos os softwares comparados, a variação de parâmetros testada no Vetorial foi $K \in [2 : n]$ com *step size* de 1.

Os métodos Vetorial, MCL e TransClust apresentaram um desempenho significativamente superior aos demais e equivalentes entre si (considerando teste de χ^2 com 95% de confiança entre as métricas F1-Score, precisão e sensibilidade). A principal diferença na composição do F1-Score é o valor de precisão obtido por eles. O método Vetorial obteve o maior valor de precisão e portanto apresenta entre 5 e 6% menos falsos positivos que os métodos MCL e TransClust.

Os métodos para clusterização do CD-HIT e USEARCH apresentaram um alto valor de precisão e um baixo valor de sensibilidade indicando a presença de falsos negativos entre 42 e 49% em comparação com os métodos Vetorial, MCL e TransClust. Este comportamento é consequência da definição de um grau arbitrário de similaridade entre alinhamento entre as sequências de aminoácidos como critério para clusterização. A proposta destes softwares é utilizar este critério e priorizar a velocidade para viabilizar clusterizações em bases de dados grandes (*Big Data*), portanto os resultados não enfraquecem seus objetivos (que não é busca de homologia). Porém é comum observar em publicações o uso destes softwares em conjuntos de dados relativamente pequenos onde seria viável o uso de métodos mais robustos, os resultados sugerem que nestes casos o uso de métodos mais adequados pode melhorar significativamente a qualidade dos grupos (teste de χ^2 com 95% de confiança) e gerar quase 50% menos falsos positivos.

A proposta dos softwares NCBI BLAST e PHMMER é encontrar proteínas com alta probabilidade de serem homólogas priorizando a precisão. Os resultados refletem este objetivo e justificam o propósito de métodos como o MCL e TransClust, que utilizam os dados gerados pelo NCBI BLAST ou PHMMER para melhorar a sensibilidade de seus resultados e identificar grupos de proteínas homólogas. Os resultados sugerem que utilizar apenas os melhores hits destes softwares implica que as sequências encontradas tem grande chance de serem homólogas porém a família da proteína estará fragmentada devido à baixa sensibilidade (0.04). Estes resultados corroboram com o conhecimento prévio que se tem em relação às limitações presentes nestes métodos (DURBIN et al., 2002).

Método	F1-Score	Precisão	Sensibilidade
Vetorial	0.97	0.97	0.98
MCL	0.95	0.92	0.98
TransClust	0.91	0.85	0.99
CD-HIT	0.72	0.99	0.56
USEARCH	0.66	0.99	0.49
PHMMER	0.08	0.92	0.04
NCBI BLASTP	0.07	0.90	0.04

Tabela 4: Melhores resultados dos métodos para agrupamento executados na base de dados *GOLD*.

2.4 PROTOCOLO DE AVALIAÇÃO: MÉTRICAS INTERNAS PARA DETERMINAR O MELHOR AGRUPAMENTO

Aplicações reais de métodos de clusterização frequentemente não permitem o uso de métricas externas pois não se conhece *a priori* as classififcações dos membros dos conjunto. Este capítulo apresenta os critérios e protocolos utilizados para comparar e avaliar as principais métricas internas que auxiliam a determinar os melhores parâmetros para agrupamento de forma não subjetiva.

2.4.1 SELEÇÃO DAS MÉTRICAS INTERNAS

A Tabela 5 representa as principais métricas internas utilizadas para clusterizações de dados biomédicos. O critério utilizado para determiná-las como principais foi sua aplicação em publicações em revistas de alto impacto (WIWIE C.; BAUMBACH; RöTTGER, 2015).

Métrica	Melhor se...	Range
Calinski-Harabasz	Maior	$[-\infty, +\infty]$
Davies-Bouldin	Menor	$[-\infty, +\infty]$
Dunn Index	Maior	$[-\infty, +\infty]$
Silhouette	Maior	$[-1, 1]$

Tabela 5: Principais métricas internas para avaliação de agrupamentos de dados biomédicos.

2.4.2 ESTUDO DE CASO 1: BASE DE DADOS DE PROTEÍNAS HOMÓLOGAS

A base de dados *GOLD* (BROWN et al., 2006) é padrão-ouro para comparações de inferência de homologia (BERNARDES J. S.; VIEIRA; ZAVERUCHA, 2015)(ROTTGER R.; KALAGHATGI; BAUMBACH, 2013) e suas classificações em nível de família de proteínas foram consideradas como referência para o cálculo das métricas externas F1-Score, precisão e sensibilidade.

2.4.3 ESTUDO DE CASO 2: BASE DE DADOS ÍRIS

Avaliar a performance das métricas internas somente em uma base de dados de proteínas homólogas pode ocultar uma eventual falta de capacidade de generalização e criar resultados que estão com forte viés para o conjunto observado (*overfitting*).

Para minimizar este efeito testes adicionais foram executados na base de dados *Íris* (FISHER et al., 1936), clássica na área de aprendizagem de máquina e que não possui nenhuma similaridade entre a de proteínas homólogas avaliada no Estudo de Caso 1.

2.4.4 FLUXOGRAMA DO PROTOCOLO DE TESTES: MÉTRICAS INTERNAS

A Figura 11 possui o fluxograma do protocolo utilizado para efetuar os testes comparativos entre as métricas internas para determinar o melhor de agrupamentonas bases de dados (WIWIE C.; BAUMBACH; RöTTGER, 2015). Na base de dados *GOLD* o método Vetorial proposto nesta dissertação foi utilizado para efetuar os agrupamentos. Na base de dados *Íris* os métodos tradicionais K-Means, K-Medoids e Linkage Average foram utilizados para efetuar os agrupamentos.

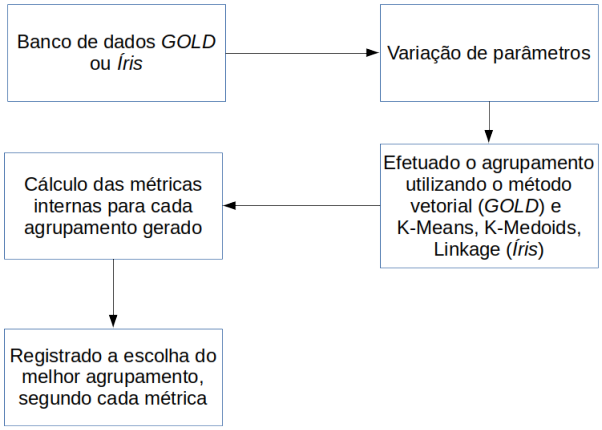


Figura 11: Fluxograma do protocolo de testes das métricas internas.

2.5 RESULTADOS COMPLEMENTARES: NEURAL NETWORK INDEX (NNI)

A Tabela 6 possui os resultados obtidos na execução do protocolo de testes descrito nas seções anteriores na base de dados *GOLD* e a Tabela 7 os resultados obtidos na base de dados *Íris*. A métrica interna Neural Network Index (NNI) foi desenvolvida neste trabalho e está definida na seção 2.1 desta dissertação, o protocolo de testes aplicado foi o mesmo para todas as métricas comparadas.

A métrica NNI apresentou o melhor desempenho entre os resultados observados (considerando teste de χ^2 com 95% de confiança entre as métricas externas obtidas) nos dois estudos de caso, demonstrando que foi significativamente superior às outras para determinar a melhor configuração de parâmetros para os métodos de clusterização testados. O desempenho superior utilizando métodos para clusterização diferentes aplicados em bases de dados distintas sugere que a métrica NNI tem capacidade de generalização e tem baixa probabilidade de estar em *overfitting* com o tipo de conjunto de dados.

Métrica	**F1-Score**	**Precisão**	**Sensibilidade**
NNI	0.91	0.87	0.95
Davies Bouldin	0.09	0.04	0.99
Dunn	0.16	0.08	0.99
Silhouette	0.10	0.05	0.99
Calinski Harabasz	0.41	0.95	0.26

Tabela 6: Resultados da avaliação dos agrupamentos selecionados a partir das métricas internas na base de dados *GOLD*

Métrica	F1-Score	Algoritmo de agrupamento	Melhor F1-Score possível
NNI	0.81	Linkage Average	0.84
Davies Bouldin	0.00	Linkage Average	0.84
Dunn	0.01	Linkage Average	0.84
Silhouette	0.00	Linkage Average	0.84
Calinski Harabasz	0.00	Linkage Average	0.84
NNI	0.82	K-Means	0.82
Davies Bouldin	0.00	K-Means	0.82
Dunn	0.01	K-Means	0.82
Silhouette	0.00	K-Means	0.82
Calinski Harabasz	0.00	K-Means	0.82
NNI	0.73	K-Medoids	0.83
Davies Bouldin	0.00	K-medoids	0.83
Dunn	0.01	K-Medoids	0.83
Silhouette	0.00	K-Medoids	0.83
Calinski Harabasz	0.00	K-Medoids	0.83

Tabela 7: Resultados da avaliação dos agrupamentos selecionados a partir das métricas internas na base de dados *Íris*. A coluna 'Melhor F1-Score possível' se refere ao maior valor de F1-Score observado entre todas as variações de parâmetros testadas ao se desconsiderar a seleção a partir da métrica interna, ou seja, é o maior valor potencial do método de agrupamento utilizado (processo análogo ao da seção 2.2.3).

3 CONCLUSÃO

O método vetorial criado para agrupar proteínas por inferência de homologia apresentou o maior F1-Score em relação às ferramentas avaliadas e a maior estabilidade entre precisão e sensibilidade.

A métrica NNI criada foi capaz de selecionar um resultado de agrupamento próximo ao melhor possível nos testes executados na base de dados de proteínas homólogas. Superando as limitações do atual estado da arte.

A NNI também superou as métricas do estado da arte em um problema tradicional (base de dados *Íris*) na determinação da melhor conformação de clusters.

REFERÊNCIAS

ANDREOPOULOS B., A. A. W. X.; SCHROEDER, M. A roadmap of clustering algorithms: finding a match for a biomedical application. **BRIEFINGS IN BIOINFORMATICS**, v. 10, p. No 3, 297–314, 2009.

BERNARDES J. S.; VIEIRA, F. R. J. C. L. M. M.; ZAVERUCHA, G. Evaluation and improvements of clustering algorithms for detecting remote homologous protein families. **BMC Bioinformatics**, v. 2015, p. 16:34, 2015.

BROWN, S. D. et al. A gold standard set of mechanistically diverse enzyme superfamilies. **Genome Biology**, v. 7, p. Issue I, Article R8, 2006.

CAMACHO C.; COULOURIS, G. A. V. M. N. P. J. B. K.; MADDEN, T. L. Blast+: architecture and applications. **BMC Bioinformatics**, v. 10:421, 2009.

CHEN, F. et al. Orthomcl-db: querying a comprehensive multi-species collection of ortholog groups. **Nucleic Acids Research**, Vol. 34, p. D363–D368, 2006.

CHEN, J. et al. A comprehensive review and comparison of different computational methods for protein remote homology detection. **Briefings in Bioinformatics**, v. 19(2), p. 231–244, 2018.

CONSORTIUM, T. U. Uniprot: the universal protein knowledgebase. **Nucleic Acids Research**, VOL. 45, p. D158–D169, 2016.

DURBIN, R. et al. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. **Cambridge University Press**, 2002.

EDGAR, R. C. Search and clustering orders of magnitude faster than blast. **Bioinformatics**, Vol. 26, No. 19, p. 2460–2461, 2010.

ENRIGHT A. J.; DONGENAND, S. V.; OUZOUNIS, C. A. An efficient algorithm for large-scale detection of protein families. **Nucleic Acids Research**, VOL. 30 NO. 7, p. 1575–1584, 2002.

FINN, R. D. et al. The pfam protein families database: towards a more sustainable future. **Nucleic Acids Research**, Vol. 44, p. D279–D285, 2015.

FISHER, R. A.; S.D.; F.R.S. The use of multiple measurements in taxonomic problems. **Annals of Eugenics**, 1936.

HAFT, D. H. et al. Tigrfams and genome properties in 2013. **Nucleic Acids Research**, Vol. 41,, p. D387–D395, 2012.

HASSANI, M.; SEIDL, T. Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. **Vietnam Journal of Computer Science**, v. 4, p. 171–183, 2016.

KRISTENSEN, D. M. et al. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. **Bioinformatics**, Vol. 26, No. 12, p. 1481–1487, 2010.

KRIVENTSEVA, E. V. et al. Orthodb v8: update of the hierarchical catalog of orthologs and the underlying free software. **Nucleic Acids Research**, Vol. 43, p. D250–D256, 2014.

LETUNIC, I.; BORK, P. 20 years of the smart protein domain annotation resource. **Nucleic Acids Research**, Vol. 46, p. D493–D496, 2017.

LEVER J.; KRZYWINSKI, M.; ALTMAN, N. Classification evaluation. **Nature Methods**, v. 13 No. 8, 2016.

LI, W.; GODZIK, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. **Bioinformatics**, Vo. 22, No 13, p. 1658–1659, 2006.

LIU, Y. et al. Understanding of internal clustering validation measures. **IEEE International Conference on Data Mining**, v. 10, p. 1550–4786, 2010.

MARCHLER-BAUER, A. et al. Cdd/sparcle: functional classification of proteins via subfamily domain architectures. **Nucleic Acids Research**, Vol. 45, p. D200–D203, 2016.

MISTRY J.; FINN, R. D. E. S. R. B. A.; PUNTA, M. Challenges in homology search: Hmmer3 and convergent evolution of coiled-coil regions. **Nucleic Acids Research**, v. 1 - 10, 2013.

PARK, H.-S.; JUN, C.-H. A simple and fast algorithm for k-medoids clustering. **Expert Systems with Applications**, Vol. 36, p. 3336–3341, 2009.

ROTTGER R.; KALAGHATGI, P. S. P. S. S. C. A. V. W. T.; BAUMBACH, J. Density parameter estimation for finding clusters of homologous proteins—tracing actinobacterial pathogenicity lifestyles. **Bioinformatics**, v. 29, p. No 2, 215–222, 2013.

SARIPELLA G. V.; SONNHAMMER, E. L. L.; FORSLUND, K. Benchmarking the next generation of homology inference tools. **Bioinformatics**, v. 32(17), p. 2636–2641, 2016.

SIGRIST, C. J. A. et al. Prosite: A documented database using patterns and profiles as motif descriptors. **Briefings in Bioinformatics**, Vol. 3, No. 3, p. 265–274, 2002.

SUZEK, B. E. et al. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. **Bioinformatics**, v. 31(6, p. 926–932, 2015.

VU D.; SZöKE, S. W. C. B. J. C. G. R. R.; ROBERT, V. Massive fungal biodiversity data re-annotation with multi-level clustering. **Nature Scientific Reports**, v. 4:6837, 2014.

WIWIE C.; BAUMBACH, J.; RöTTGER, R. Comparing the performance of biomedical clustering methods. **Nature Methods**, v. 12, p. 11, 2015.

XU, R.; WUNSCH, D. C. Clustering algorithms in biomedical research: A review. **IEEE REVIEWS IN BIOMEDICAL ENGINEERING**, v. 3, p. 1937–3333, 2010.

ANEXO A – AVALIAÇÃO DAS PROJEÇÕES

O método vetorial proposto possui um processo de redução de dimensionalidade (projeção da matriz em uma base ortonormal) que causa perda de informações. A Figura 12 demonstra a relação entre o F1-Score máximo obtido em cada um dos tamanhos de projeções avaliadas na base de dados *GOLD*. Para o estudo de caso utilizamos a projeção de tamanho 5500 que apresentou um F1-Score de 0.97.

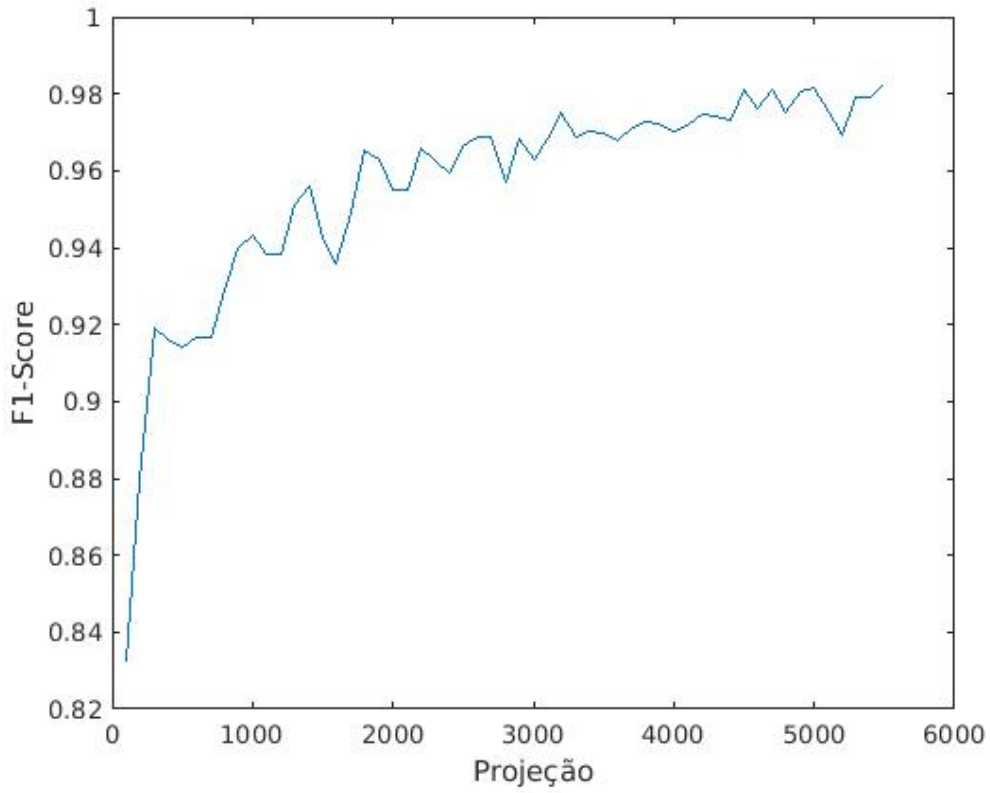


Figura 12: Testes de projeções em relação ao F1-Score obtido na base de dados *GOLD*.

ANEXO B – AVALIAÇÃO DE ALGORITMOS PARA CLUSTERIZAÇÃO

Após a extração dos atributos (representação vetorial de sequências de aminoácidos nesta dissertação) é necessário aplicar um método para realizar a clusterização efetivamente. Para avaliar o potencial dos algoritmos disponíveis para identificação de grupos homólogos foram testados os métodos da Tabela 8 executando todas as variações possíveis de seus parâmetros (por exemplo K-Means $K \in [2, n]$, sendo n o número total de elementos no conjunto) e todas as distâncias disponíveis. Para diminuir o tamanho da tabela, apenas o melhor valor de F1-Score obtido pela melhor distância está sendo apresentado.

Algoritmo	**Maior F1-Score**
linkage average (Matlab 2017a)	0.9796
hcclust average (R stats 3.4.3)	0.9769
hcclust single (R stats 3.4.3)	0.9760
hcclust complete (R stats 3.4.3)	0.9436
hcclust median (R stats 3.4.3)	0.7377
Kmeans Lloyd's (Matlab 2017a)	0.6464
hcclust centroid (R stats 3.4.3)	0.5367
Kmedoids (Matlab 2017a)	0.4518
kmeans Hartigan-Wong (R stats 3.4.3)	0.4050
Kohonen (Matlab 2017a)	0.3668
hcclust ward2 (R stats 3.4.3)	0.3341
Kohonen (R kohonen v3.0.4)	0.2735
hcclust ward (R stats 3.4.3)	0.2582
dbscan (R fpc v2.1-11)	0.2296
clara (R cluster v2.0.6)	0.1572
pam (R cluster v2.0.6)	0.1504
CrossClustering (R CrossClustering v3.0)	0.04
Fanny (R cluster v2.0.5)	0.00

Tabela 8: Algoritmos para clusterização avaliados e os maiores valores de F1-Score obtidos na base de dados *GOLD*.