Alunos: Gustavo de Souza e José Augusto Laube

Trabalho final complexidade de algoritmos

Introdução:

A Bioinformática emerge como uma disciplina essencial na era pós-genômica, utilizando o poder computacional para decifrar a complexa teia de interações que governam a vida. Uma das abordagens mais poderosas neste campo é a modelagem de sistemas biológicos como redes, onde as Redes de Interação de Proteínas (PPIs) representam as relações funcionais em uma célula. Muitos problemas fundamentais na análise dessas redes, como a identificação de proteínas-chave, são computacionalmente intratáveis (NP-difíceis), tornando inviável encontrar soluções ótimas para redes de grande escala. Diante deste desafio, o uso de heurísticas em tempo polinomial se torna uma estratégia indispensável. Este trabalho explora a aplicação de uma heurística gulosa para o Problema da Cobertura Mínima de Vértices, com o objetivo de identificar um subconjunto de proteínas estruturalmente críticas na rede de levedura, revelando insights sobre a organização e a robustez dos sistemas biológicos.

Definição do problema:

Células são como uma fábrica muito complexa, que tem vários componentes. As proteínas são as máquinas e os trabalhadores dessa “fábrica”, em que nenhum trabalhador ou equipamento trabalha sozinho, desta forma essas proteínas se ativam, e desativam e montam estruturas em conjunto para realizar todas as tarefas necessárias para a vida. Uma interação de proteína-proteína (PPI) é uma dessas colaborações, podendo ser uma conexão física ou funcional entre duas proteínas. Ao mapear todas as proteínas de um organismo e todas as interações entre elas, obtendo uma rede de interação de proteínas, uma célula tem milhares de proteínas e dezenas de milhares de interações, essas estruturas não são aleatórias e tem suas próprias organizações. Desta forma a estrutura da rede determina como a célula funciona, como responde a mudanças e o que acontece quando algo dá errado.

Desta forma a identificação de proteínas essenciais nessa rede pode ser muito útil, principalmente quando estamos desenvolvendo algum medicamento para patógenos como bactérias ou um fungo, pois se atacarmos pontos críticos nessa rede de proteínas conseguimos causar o máximo de dano nesse organismo.

Desta forma usar uma heurística como cobertura mínima de vértices nos ajuda a identificar quais são as proteínas são essenciais para a sobrevivência desse organismo, desta forma podemos desenvolver algum medicamento que consiga remover várias dessas proteínas o que seria letal para o organismo.

Então para esse trabalho escolhemos uma rede de proteínas no site https://networkrepository.com/PROTEINS-full.php que conta com mais de 43 mil proteínas, e 162 mil conexões, nosso objetivo aqui é aplicar a heurística de conjunto mínimo de vértices, para encontrar o conjunto essencial de proteínas dessa rede.

Solução proposta:

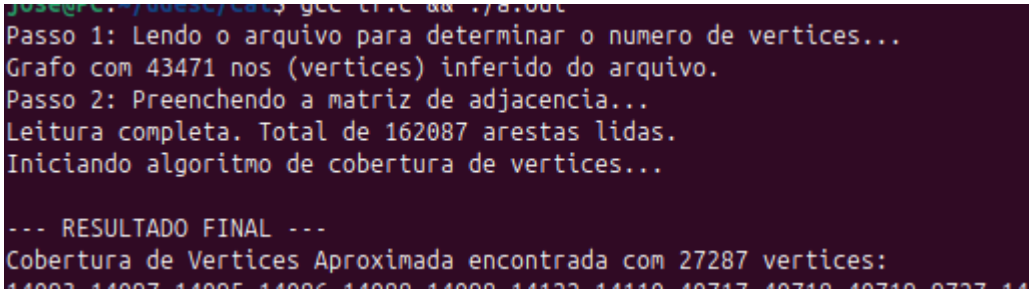
Para resolver esse problema desenvolvemos um programa que lê um arquivo csv que representa as conexões da nossa rede de proteínas, as proteínas são nomeadas de 1 até 43470,

e colocamos em uma matriz de 43470x43470 em que as linhas representam cada uma das proteínas e as colunas representam se aquela proteína tem uma conexão com a proteína da linha que está sendo analisada. Desta forma é possível contar quantas conexões têm cada um das proteínas, para não ter que contar a todo momento essa matriz, criamos um vetor de 43470 posições, sendo que o valor da posição é a quantidade de conexões que aquela proteína tem.
A lógica base para o nosso algoritmo é buscar o vértice com mais ligações nesse vetor, salvar ele em um vetor de vértices essenciais, e remover todas as suas conexões, decremento em cada uma das proteínas que ele está relacionado. O programa faz isso repetidamente até que toda a rede seja coberta, ou seja, quando não houver mais nenhuma conexão entre as proteínas

```c
void contaGrau(int **matLigacoes, int *vetMaiorGrau, int numVertices) {
    for(int i = 0; i < numVertices; i++){
        int total = 0;
        for(int j = 0; j < numVertices; j++){
            if(matLigacoes[i][j] == 1){
                total++;
            }
        }
        vetMaiorGrau[i] = total;
    }
}

int maiorGrau(int *vetMaiorGrau, int numVertices) {
    int maior = -1;
    int posi = 0;
    for(int i = 0; i < numVertices; i++){
        if(vetMaiorGrau[i] > maior){
            maior = vetMaiorGrau[i];
            posi = i;
        }
    }
    return posi;
}

void removeGrau(int **matLigacoes, int *vetMaiorGrau, int maiorVert, int numVertices) {
    for(int i = 0; i < numVertices; i++){
        if(matLigacoes[i][maiorVert] == 1){
            matLigacoes[i][maiorVert] = 0;
            matLigacoes[maiorVert][i] = 0;
            vetMaiorGrau[i]--;
        }
    }
    vetMaiorGrau[maiorVert] = -1;
}

void coberturaDeVertices(int **matLigacoes, int *vetMaiorGrau, int **vertices, int *tam, int numVertices) {
    int maiorVert = maiorGrau(vetMaiorGrau, numVertices);

    if (vetMaiorGrau[maiorVert] <= 0) {
        return;
    }

    (*tam)++;
    *vertices = realloc(*vertices, sizeof(int) * (*tam));
    if (*vertices == NULL) {
        perror("Erro ao realocar memoria para vertices");
        exit(1);
    }
    (*vertices)[(*tam)-1] = maiorVert + 1;

    removeGrau(matLigacoes, vetMaiorGrau, maiorVert, numVertices);
}
```

Testes e resultados:
Como a rede escolhida para fazer essa análise é muito grande, testamos primeiro com uma rede menor de 2000 proteínas, o que era um pouco mais rápido de se resolver, porém com redes muito grandes como a escolhida para esse trabalho, os resultados demoram um pouco.

Com o código final concluído conseguimos identificar a rede central de interação de proteínas desse organismo, foram encontradas 27287 proteínas nessa rede que são consideradas essenciais, sendo assim cerca de 67.77% das proteínas dessa rede são essenciais para o funcionamento completo desta rede.



O tamanho considerável da cobertura obtida desafia a hipótese simplista de que a integridade da rede é mantida por um pequeno núcleo de proteínas "super essenciais". Em vez disso, o resultado sugere que a arquitetura da rede de interações de proteínas é altamente distribuída e interligada. Não há um pequeno conjunto de "calcanhares de Aquiles" cuja remoção seria suficiente para fragmentar toda a rede. Pelo contrário, a funcionalidade celular parece depender da participação integrada de uma proporção massiva de suas proteínas.
Este achado aponta para uma propriedade biológica fundamental: a robustez. Uma rede que exige que mais de 60% de seus nós sejam removidos para que todas as suas conexões diretas cessem é inerentemente resiliente a perturbações. Isso faz sentido do ponto de vista evolutivo, pois a falha de algumas poucas proteínas (devido a mutações ou estresse ambiental) não levaria a um colapso catastrófico do sistema. A grande cobertura sugere um alto grau de redundância funcional e de interconexão, onde múltiplas proteínas e vias podem compensar falhas umas das outras.

Conclusão:
A heurística gulosa não garante a solução mínima ótima. É possível que, ao fazer escolhas localmente ótima, o algoritmo tenha sido forçado a selecionar mais nós do que o estritamente necessário em comparação com uma solução ótima. O tamanho real da cobertura mínima é provavelmente menor, mas o fato de uma heurística poderosa ainda exigir um conjunto tão grande é, por si só, significativo.
O resultado é também um reflexo direto da topologia da rede analisada. Ele sugere que a rede pode não ser dominada por alguns poucos "mega-hubs", mas sim por uma distribuição mais ampla de "hubs" de tamanho médio e uma grande quantidade de interações entre nós de baixo grau, forçando o algoritmo a "caçar" e adicionar muitos nós para cobrir todas as arestas "periféricas".
Em conclusão, este experimento demonstra que a aplicação da Cobertura Mínima de Vértices em redes de interação de proteínas não apenas identifica um conjunto de nós estruturalmente importantes, mas também revela características globais da arquitetura celular. O resultado de 62.8% sugere que a robustez e a complexidade funcional das células vivas são sustentadas

por um sistema vasto e descentralizado, em vez de um pequeno núcleo de controle, fornecendo uma visão quantitativa da resiliência inerente à vida.

Referências:

SERPA, Matheus S.; RODRIGUES, Thiago N.; ALVES, Ítalo C.; et al. Análise de Algoritmos. Porto Alegre: SAGAH, 2021. *E-book.* p.252. ISBN 9786556901862. Disponível em: https://app.minhabiblioteca.com.br/reader/books/9786556901862/. Acesso em: 07 jul. 2025.

Balaji, S., Swaminathan, V. e Kannan, K. (2010). An empirical study of heuristics for the vertex cover problem. *International Journal of Computer Theory and Engineering*, v. 2, n. 2, p. 1793-8201.

Barabási, A.-L. e Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, v. 5, n. 2, p. 101-113.