MAC0215 - Proposta

Aluno: Igor Fratel Santana
Orientador: Alan Mitchell Durham

2º Semestre de 2017

1 Resumo

A identificação de proteínas homólogas, que compartilham um ancestral comum, é essencial para obtermos uma melhor precisão na predição da função de cada proteína e também tem um papel importante em genômica comparativa (estudo da relação entre genomas de diferentes espécies ou linhagens biológicas, desde estruturas às funções dos mesmos). Para este fim, foram criados programas identificadores de pares de proteínas homólogas (nesse trabalho, também usaremos métodos para detecção de pares ortólogos, separados por eventos de especiação), que em geral utilizam a similaridade entre duas sequências de aminoácidos, uma comparação entre os domínios de cada proteína ou a cobertura do alinhamento entre duas sequências de aminoácidos para calcular um *score* para os pares de proteínas a partir do qual podemos avaliar a possibilidade de homologia.
Temos como objetivo aplicar métodos de detecção de homologia e ortologia para o agrupamento de vizinhanças gênicas (genes vizinhos no mesmo cromossomo, um ao lado do outro).
Este projeto faz parte da minha iniciação científica com o professor Alan Mitchell Durham.

2 Introdução

Nas próximas subseções serão apresentadas as proteínas, com foco nas informações utilizadas pelos identificadores de homologia, e as técnicas usadas em bioinformática para inferir homologia e ortologia. Em seguida, exporemos brevemente um método já existente para identificação de homologia e ortologia que será incorporado neste trabalho. Por último, informarei a situação atual do projeto.

2.1 Proteínas e homologia

Proteínas são macromoléculas biológicas constituídas por uma ou mais cadeias de aminoácidos. As proteínas estão presentes em todos os seres vivos e participam em praticamente todos os processos celulares, desempenhando um vasto conjunto de funções no organismo, como a replicação de DNA, a resposta a estímulos e o transporte de moléculas.
As proteínas podem possuir domínios, que são partes da cadeia de aminoácidos que podem assumir uma estrutura tridimensional independente e estável. A existência dos domínios permite a construção de proteínas a partir de módulos funcionais, como mostrado na figura 1.

O termo homologia aqui empregado parte do contexto da genética, aplicado a genes e seus produtos (aqui usaremos as proteínas), significando sequências que possuem um ancestral comum.
A identificação de proteínas homólogas é essencial para obtermos uma melhor precisão na predição da função de cada proteína e também tem um papel importante em genômica comparativa (estudo da relação entre genomas de diferentes espécies ou linhagens biológicas, desde estruturas às funções dos mesmos).
Os métodos de detecção de homologia/ortologia aqui usados serão os *graph-based*, de acordo com o *review* "*The quest for orthologs: finding the corresponding gene across genomes*"(https://www.ncbi.nlm.nih.gov/pubmed/18819722). Esses métodos se baseiam na similaridade de sequência entre todas as sequências envolvidas. Essa similaridade pode ser calculada usando diversos métodos de alinhamento de pares de sequências (algoritmos que posicionam duas sequências paralelamente de forma a encontrar regiões de similaridade), como o BLAST (basic local alignment search tool) ou o algoritmo de alinhamento local Smith-Waterman.
A figura 2 mostra as formas básicas de agrupar proteínas utilizando as medidas discutidas acima.

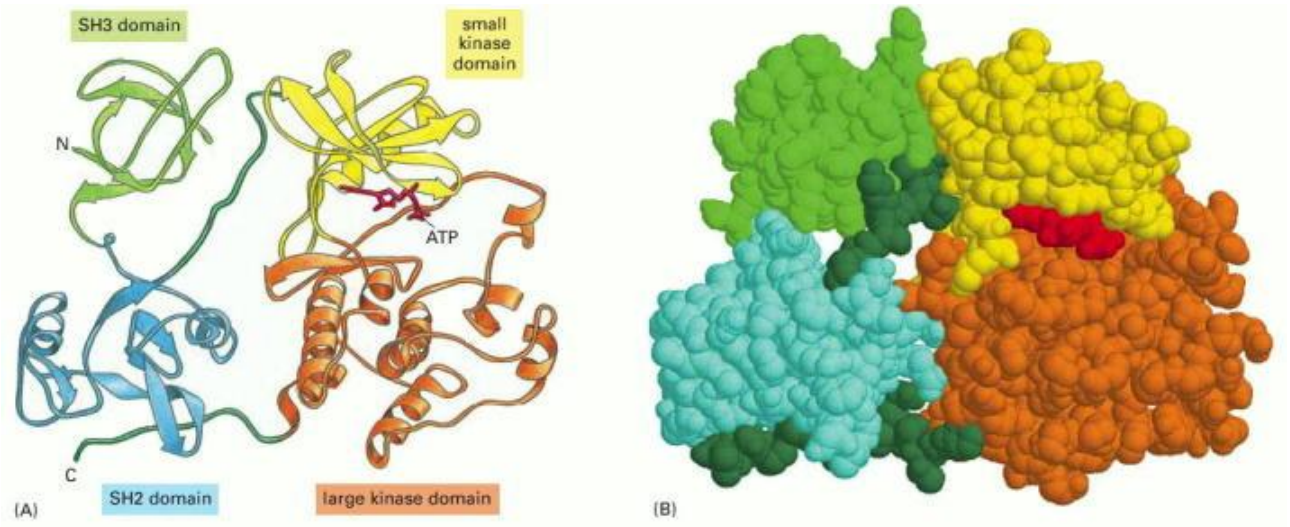

Figura 1: *Exemplo de uma proteína que possui quatro domínios*

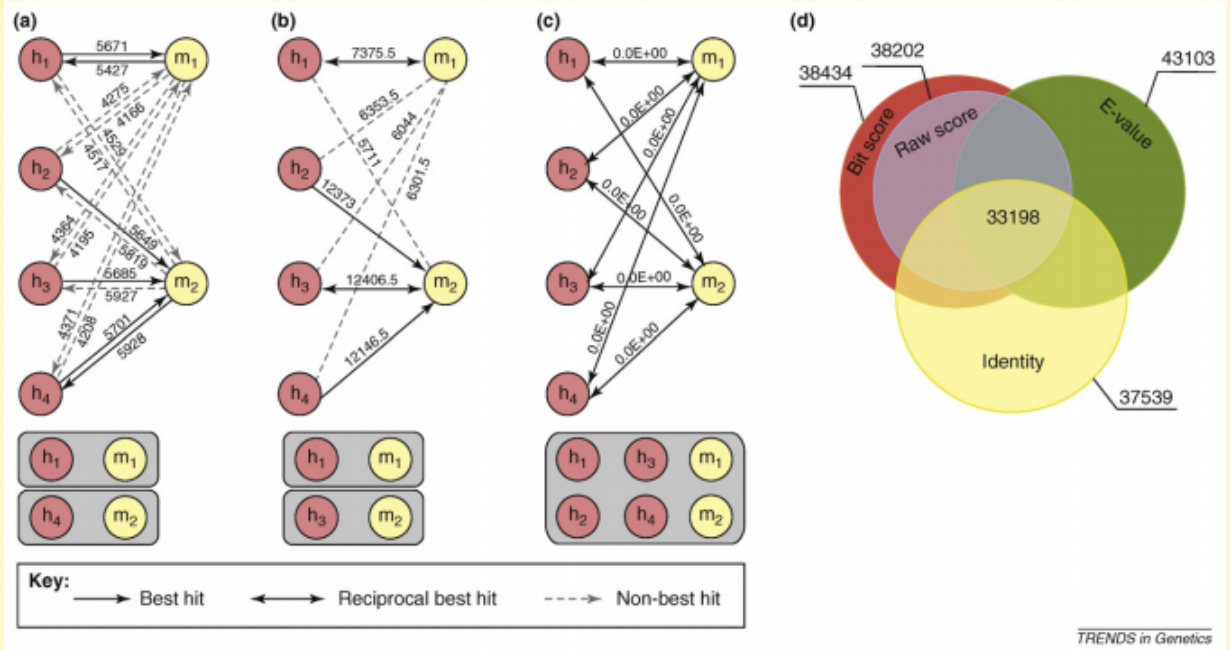

Figure I. Different sets of putative orthologs defined as reciprocal best hits. Three graphs of human (h_1–h_4) and mouse (m_1- and m_2) mucin-5 proteins are constructed using three different protein similarity measures: **(a)** asymmetric BLAST raw score; **(b)** symmetric Smith-Waterman score; and **(c)** symmetric BLAST E-value. The corresponding set of predicted orthologs is shown below each graph. Clearly, the reciprocal best hit approach using different similarity measures might result in different but largely overlapping sets of orthologs. **(d)** Venn diagram of four different sets of orthologs, using BLAST identity, E-value, raw and bit score, are inferred from complete human and mouse proteomes (Refseq version 29). The total number of orthologs is indicated for the sets and four-way intersection. Graph nodes correspond to RefSeq protein accessions: h_1, XP_001717932; h_2, NP_059981 (*Muc5ac*); h_3, NP_002449 (*Muc5b*); h_4, XP_001719401; m_1, NP_034974 (*Muc5ac*); m_2, NP_083077 (*Muc5b*).

Figura 2

No contexto do projeto, trataremos as proteínas como *strings* de aminoácidos ou como *strings* de domínios, dependendo do método usado.

2.2 Neighborhood Correlation

Nesse projeto pretendemos incorporar o programa Neighborhood Correlation (http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000063), que fornece uma forma de calcular um *score* para pares de proteína, se valendo da estrutura das redes de similaridade de proteínas.
Redes de similaridade são grafos onde as proteínas são vértices e as arestas indicam um nível de similaridade entre as proteínas.
Basicamente, o Neighborhood Correlation calcula um *score* para um par de proteínas baseado nas vizinhanças únicas e compartilhadas de cada proteínas do par. Nesse caso, a vizinhança de uma proteína significa o conjunto de proteínas ligadas a ela na rede de similaridade e, quanto maior a vizinhança compartilhada em relação à vizinhança única de cada proteína, mais provável é a homologia entre elas.

2.3 Situação atual do projeto

Nos últimos meses eu li o artigo que apresenta o método Neighborhood Correlation, chamado "Sequence Similarity Network Reveals Common Ancestry of Multidomain Proteins"(http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000063 e outro artigo, apresentando o programa porthoDom (https://www.ncbi.nlm.nih.gov/pubmed/25968113), que não foi incluído na proposta por ser insatisfatório em uma primeira análise.
Eu também comecei a análise da qualidade do método Neighborhood Correlation, tentando reproduzir os dados apresentados no artigo. Para tal, escrevi um programa que gera as medidas e gráficos necessários. Ainda preciso otimizá-lo e executá-lo com os dados corretos.
Além disso, escrevi um programa capaz de realizar um alinhamento global de sequências arbitrárias (https://github.com/igorfratel/First_Alignment), que pode vir a ser útil ao projeto.

3 Objetivos

3.1 Definições

Dados N fragmentos de genoma com $2l+1$ genes cada, onde l é um numero natural, temos $N*(2l+1)$ genes. O gene $g_{x,y}$ é o x-ésimo gene do fragmento G_y, onde $1 \leqslant x \leqslant 2l+1$ e $1 \leqslant y \leqslant N$. Cada $g_{x,y}$ tem uma relação bijetiva com uma proteína $p_{x,y}$.

3.2 Proposta

A proposta final do projeto é, recebendo um conjunto de fragmentos de genomas de diferentes organismos com um gene "âncora"comum a todos eles (idealmente o l+1-ésimo gene, deixando l genes à esquerda e à direita), agrupar as proteínas traduzidas pelos $N*(2*l+1)$ genes usando uma medida de detecção de homologia/ortologia (inicialmente o Neighborhood Correlation).
Utilizando as classificações das proteínas em grupos, podemos adicionar uma informação nova aos genes precursores. A partir dessa nova informação sobre os genes, desejamos agrupar os fragmentos gênicos de acordo com a semelhança das suas vizinhanças (subsequência de genes em um genoma) utilizando alguma medida a ser proposta.
A sequência de passos do programa está representada de forma geral nas imagens 3 e 4. A análise da ordem dos genes em vizinhanças gênicas pode ser útil para identificar padrões de genes contíguos se repetindo em genomas evolutivamente distantes, o que pode indicar a existência de uma relação funcional entre eles.

3.3 StringDB

O StringDB (https://string-db.org/) fornece funcionalidade similares ao que pretendemos fazer no menu de visualização "neighborhoods". A principal diferença é que o StringDB não oferece flexibilidade em relação ao método de agrupamento e ao banco de dados usado.

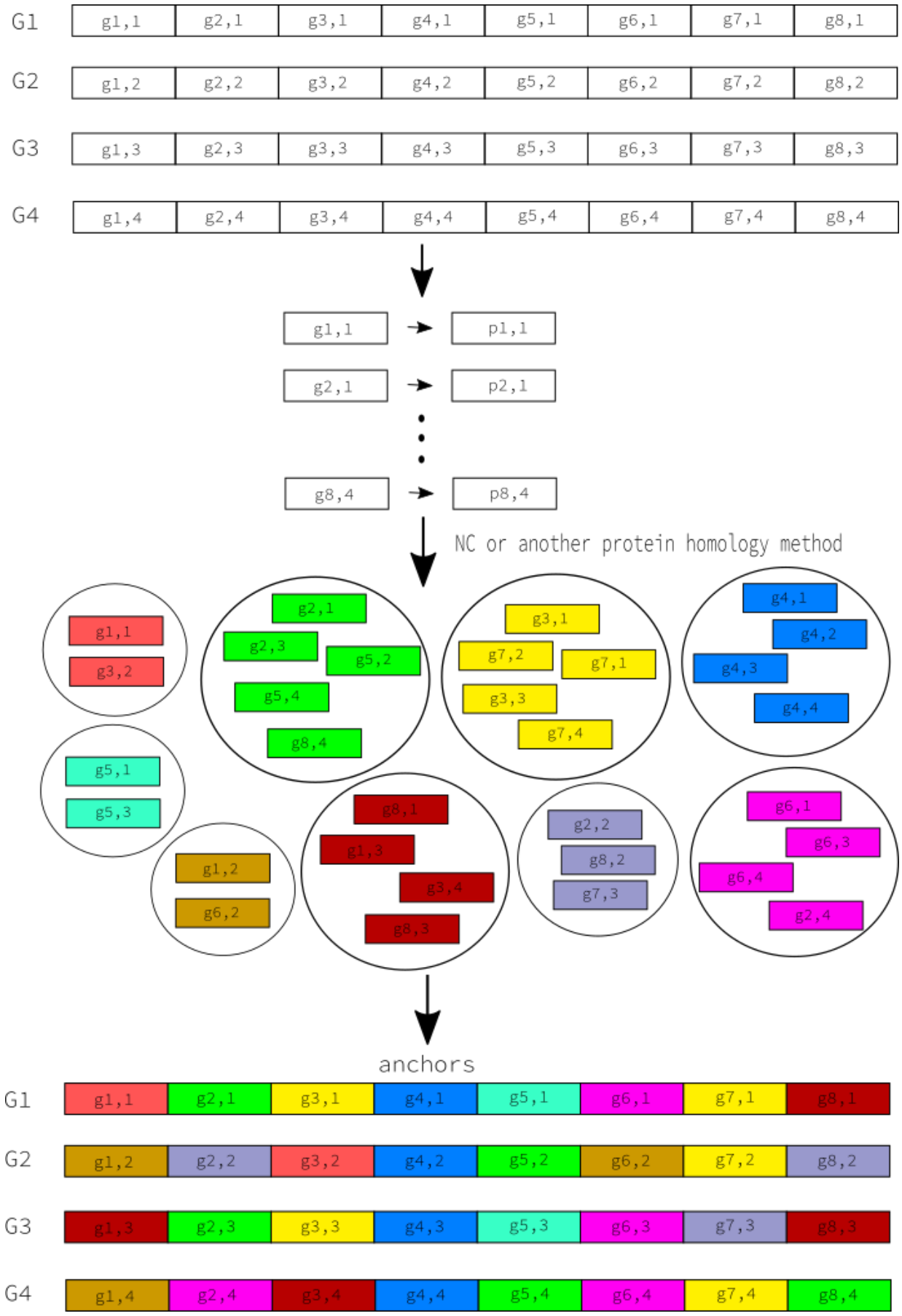

Figura 3: *Agrupamos as proteínas de acordo com algum método de detecção de homologia/ortologia e dessa forma obtemos uma nova informação sobre os genes.*

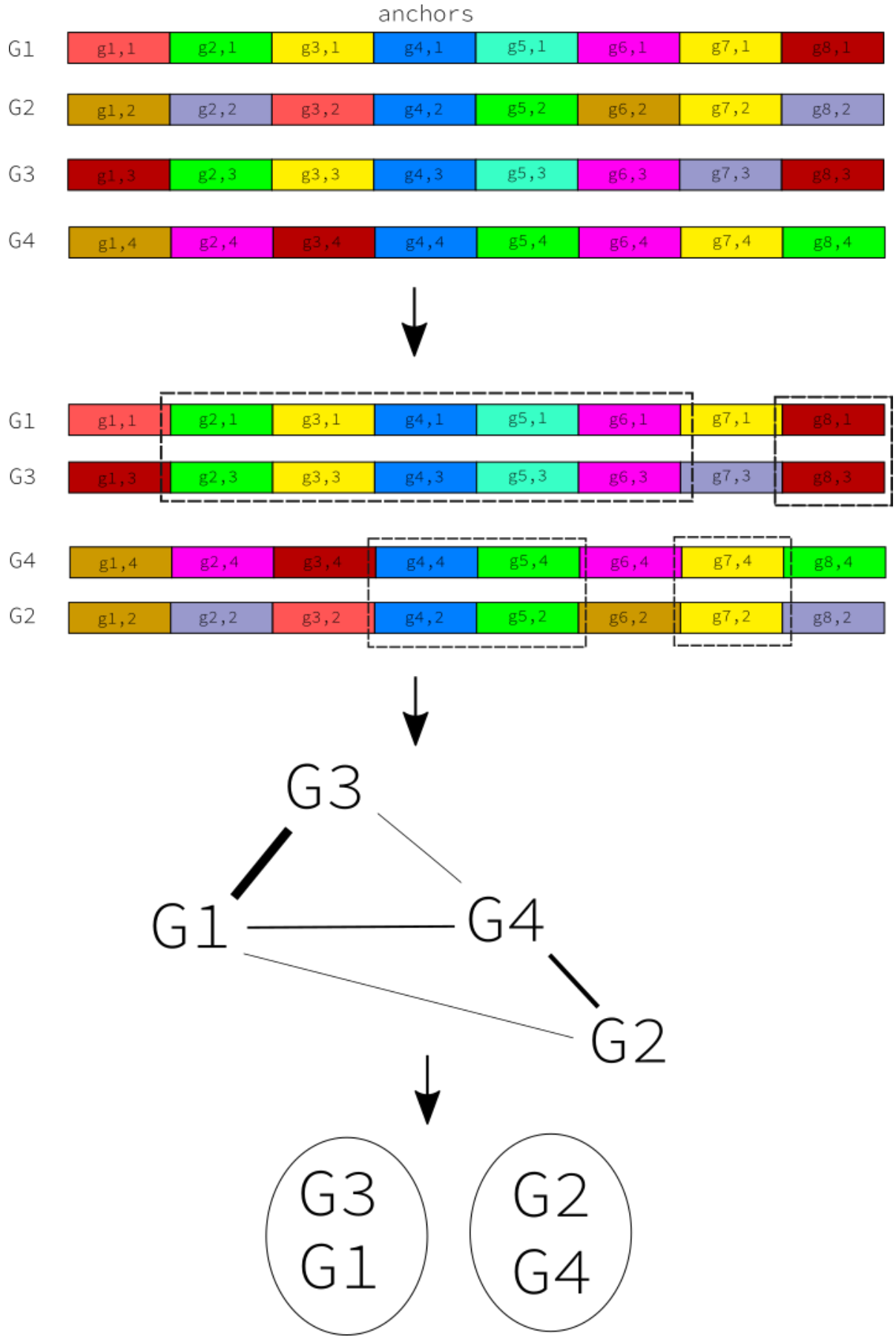

Figura 4: *A partir da nova informação obtida sobre os genes, podemos agrupá-los e visualizá-los de diferentes formas de acordo com a similaridade das suas vizinhanças. No exemplo, eles foram ordenados, transformados em grafos com a espessura das arestas indicando a similaridade e, finalmente, separados em conjuntos disjuntos.*

Nesse projeto, pretendemos deixar as escolha de métodos de agrupamento o mais flexível quanto possível, além de permitir que o usuário especifique um conjunto de proteínas próprio para ser usado como banco de dados.

4 Tarefas

- Reproduzir os dados do Neighborhood Correlation -> completo em aproximadamente 15h
- Escrever as bases para um programa que utilize medidas de detecção de homologia e ortologia para o agrupamento de vizinhanças gênicas -> 21h
- Discutir medidas para o agrupamento das vizinhanças gênicas -> 10h
- Implementar e testar o programa usando o neigborhood correlation como medida de agrupamento de proteínas e a medida discutida no item acima para o agrupamento das vizinhanças gênicas -> 16h
- Estudar as medidas usadas pelo StringDB -> 7h
- Explorar outras medidas de detecção de homologia/ortologia - 10h
- Incluir no programa as medidas discutidas -> 15h
- Finalizar todos os detalhes do programa - 6h

5 Método de acompanhamento

Na página https://igorfratel.github.io/MAC0215/, divulgarei, pelo menos uma vez por semana, os meus avanços no projeto.