Pontifícia Universidade Católica
do Rio de Janeiro

Waldir José Pereira Junior

Clusterização de poços de petróleo utilizando alinhamento de sequências baseadas em litologia

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática, do Departamento de Informática da PUC-Rio.

Orientador: Prof. Marcelo Gattass
Co-orientador: Prof. Helio Côrtes Vieira Lopes

Rio de Janeiro,
Setembro de 2021

Waldir José Pereira Junior

Clusterização de poços de petróleo utilizando alinhamento de sequências baseadas em litologia

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática da PUC-Rio. Aprovada pela Comissão Examinadora abaixo.

Prof. Marcelo Gattass
Orientador
Departamento de Informática - PUC-Rio

Prof. Helio Côrtes Vieira Lopes
Co-orientador
Departamento de Informática - PUC-Rio

Prof. Marcus Vinicius Soledade Poggi de Aragão
Departamento de Informática - PUC-Rio

Prof. Alberto Barbosa Raposo
Departamento de Informática - PUC-Rio

Rio de Janeiro, 16 de setembro de 2021

Waldir José Pereira Junior

Graduou-se em Ciência da Computação pela Universidade Federal do Rio de Janeiro (UFRJ) em 2009. Em 2018 ingressou no Programa de Pós-Graduação em Informática com ênfase em Ciência de Dados da PUC-Rio, para obtenção do título de Mestre.
Atualmente trabalha como Líder Técnico do Instituto Tecgraf – PUC-Rio, mais especificamente na área de sistemas para gerenciamento de reservatórios e construção de poços.

Ficha Catalográfica

Pereira Junior, Waldir José

Clusterização de poços de petróleo utilizando alinhamento de sequências baseadas em litologia / Waldir José Pereira Junior; orientador: Marcelo Gattass; co-orientador: Helio Côrtes Vieira Lopes – Rio de Janeiro: PUC-Rio, Departamento de Informática, 2021.

82p.

1. Dissertação (Mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui referências bibliográficas.

1. Informática – Tese. 2. Correlação. 3. Clusterização. 4. Perfil Litológico. 5. Litologia. 6. Alinhamento de Sequências. I. Gattass, Marcelo. II. Vieira Lopes, Helio Côrtes. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. IV. Título.

CDD: 004

Aos meus pais, Marize e Waldir, por me ensinarem
metade das coisas que importam.

E aos meus filhos, Bento e Caio, por
me ensinarem a outra metade.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Ao meu orientador, professor Marcelo Gattass, por ter acreditado no potencial desta pesquisa desde o início, pelo incentivo e parceria para a realização deste trabalho. E ao meu co-orientador, professor Helio Lopes, por todo apoio, principalmente durante o início deste trabalho.

Ao Instituto Tecgraf da PUC-Rio, pelos auxílios concedidos para a realização de toda a pesquisa. Aos meus superiores diretos, Marcos Marques e Leandro Nazareth, por acreditarem no meu potencial e permitirem que eu conciliasse o estudo com os afazeres profissionais.

Aos amigos dos projetos CronoWeb e CROSS, pela paciência, principalmente nos dias após noites dedicadas a esta pesquisa. E aos demais amigos do Instituto Tecgraf da PUC-Rio, pelo apoio e incentivo.

À Petrobras, pela disponibilização de dados, sem os quais não poderia ter realizado este trabalho.

A todos os professores e funcionários do Departamento de Informática da PUC-Rio, pelos ensinamentos e pela ajuda.

À minha querida mãe, Marize, por todo o suporte e apoio, sem os quais esta empreitada não teria sido possível.

Aos meus amados filhos, Bento e Caio, por trazerem alegria em diversos momentos desta jornada.

À Isabella, por todas as discussões, orientações e apoio durante a construção deste trabalho.

À minha família, por estar sempre comigo e me apoiar em todos os momentos.

A todos que acreditaram e torceram por mim. Muito obrigado.

Resumo

Pereira Junior, Waldir José; Gattass, Marcelo; Lopes, Helio Côrtes Vieira. **Clusterização de poços de petróleo utilizando alinhamento de sequências baseadas em litologia**. Rio de Janeiro, 2021. 82p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A construção de um poço de petróleo requer um planejamento extenso e antecipado. Dentre os vários objetivos deste planejamento, está a verificação da necessidade de aquisição de materiais e equipamentos para a realização das etapas da construção do poço. Tais aquisições muitas vezes envolvem contratações longas e, posteriormente, requerem um grande tempo para entrega, podendo chegar a anos. Como este planejamento é realizado em um cenário de muitas incertezas, várias técnicas, utilizando diversos tipos de dado, já foram propostas para correlacionar poços, de modo a obter antecipadamente as necessidades de materiais e equipamentos para construir um novo poço. Um desses tipos de dado é o perfil litológico, que contém os seguimentos de rochas presentes pela extensão do poço, coletados através de sensores e outros meios presentes durante a perfuração. Este perfil litológico pode ser gerado artificialmente para poços ainda não perfurados, através de dados sísmicos. Este trabalho propõe uma nova metodologia para agrupar poços de petróleo. A medida de distância será calculada com base no grau de similaridade entre poços, obtido através da aplicação de algoritmo de alinhamento de sequências, que, por sua vez, são geradas exclusivamente a partir dos perfis litológicos de tais poços. Desta forma, é possível obter poços correlatos a um determinado poço. Para validação da metodologia, foram realizados experimentos de clusterização envolvendo dados de 120 poços da costa sudeste brasileira.

Palavras-chave

Correlação; Clusterização; Perfil Litológico; Litologia; Alinhamento de Sequências.

Abstract

Pereira Junior, Waldir José; Gattass, Marcelo (Advisor); Lopes, Helio Côrtes Vieira. **Oil well clustering using lithology-based sequence alignment**. Rio de Janeiro, 2021. 82p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The construction of an oil well requires extensive and advanced planning. Among the various objectives of this planning is the verification of the need to purchase materials and equipment to carry out the stages of construction of the well. Such acquisitions often involve long contracts and, later, require a long lead-time, which can reach years. As this planning is carried out in a scenario of many uncertainties, several techniques, using different types of data, have already been proposed to correlate wells, in order to obtain in advance the material and equipment requirements to build a new well. One of these types of data is the lithological profile, which contains the rock segments present throughout the length of the well, collected through sensors and other methods present during the drilling. It is possible to generate artificial lithological profiles for not yet drilled wells, through seismic data. This work proposes a new methodology for grouping oil wells. The distance measure is based on the degree of similarity between wells, obtained by applying a sequence alignment algorithm, which, in turn, are generated exclusively from the lithological profiles of such wells. In this way, it is possible to obtain wells related to a specific well. To validate the methodology, clustering experiments involves data from 120 wells on the southeastern Brazilian coast.

Keywords

Correlation; Clustering; Lithological Profile; Lithology; Sequence alignment.

Sumário

Lista de Figuras

Lista de Tabelas

O começo de todas as ciências é o espanto de as coisas serem o que são.

(Aristóteles)

1
Introdução

A construção de um poço de petróleo requer um planejamento muito extenso e antecipado. Este planejamento possui uma etapa dedicada à análise de prontidão de materiais e equipamentos necessários à construção do poço. A análise de prontidão procura, para cada material ou equipamento, relacionar a oferta, através de dados de estoque e de contratos de fornecimento vigentes, com a demanda ao longo do tempo. A conclusão da análise de prontidão permite que a companhia decida se e quando deverá realizar novas contratações, assim como permite definir a quantidade e o tempo de contrato. Como este planejamento se inicia alguns anos antes da construção do poço e os processos de contratação, fabricação e mobilização necessitam de muitos meses, podendo chegar a anos, é extremamente importante que a definição da demanda seja realizada corretamente.

Uma maneira comumente utilizada para definir a demanda de materiais e equipamentos é a correlação do poço que está sendo planejado com outros poços já perfurados, de modo a obter um modelo de necessidades. Esta correlação é comumente realizada de forma manual por profissionais geólogos experientes. Como é uma atividade manual e muito dependente de análise humana, os resultados não são padronizados e a execução não é facilmente escalável. No intuito de resolver este problema, diversas abordagens já foram propostas na literatura para efetuar correlação entre poços de forma automática.

Uma das maneiras de descrever um poço de petróleo é através do seu perfil litológico, que cataloga todos os intervalos de rochas existentes durante toda a extensão do poço, estruturados sequencialmente. Tais informações são obtidas através de sensores que coletam dados durante a perfuração do poço de petróleo. Porém, para poços ainda não perfurados, um perfil litológico artificial pode ser gerado através da análise dos dados sísmicos da região onde o poço será perfurado.

Uma das principais maneiras de correlacionar sequências é através de algoritmos de alinhamento de sequências, presentes na Bioinformática. Tais algoritmos podem ser exatos ou heurísticos e podem gerar resultados de

alinhamentos globais, que abrangem as sequências por inteiro, ou locais, que verificam trechos internos das sequências. Muitas vezes esses algoritmos são utilizados para realizar alinhamento entre diversas sequências, mas também podem ser utilizados para alinhamento entre duas sequências, apenas.

Dentre as várias formas de se obter agrupamentos de dados estão os algoritmos de clusterização particionais, que são métodos de aprendizado de máquina não supervisionados, pertencentes à área de Ciência de Dados. Tais algoritmos se baseiam em medidas de distância, que podem ser obtidas a partir de medidas de similaridade ou de dissimilaridade, para encontrar a melhor configuração de grupos, de modo que os elementos do mesmo grupo sejam muito semelhantes entre si e muito diferentes de elementos dos demais grupos.

1.1 Contribuição desta pesquisa

Algumas consultas foram realizadas na base científica SCOPUS através da utilização de palavras-chave relacionadas ao tema deste trabalho e outros critérios de busca. Primeiramente foi realizado um agrupamento das palavras-chave e critérios de busca, como apresentado na Tabela 1 abaixo. No grupo A estão termos relacionados ao tema de correlação, que é o objetivo central deste trabalho. No grupo B estão termos relacionados à área de negócios que este trabalho utilizou como motivação, que é a litologia de poços. Já o grupo C representa o meio que este trabalho utilizou para realizar a correlação de poços de petróleo com base em litologia, que é o alinhamento de sequências. Por fim, um filtro adicional foi incluído no grupo D, para retornar apenas trabalhos relacionados à área de computação.

Grupo	Palavras-chave
A	("correlation" OR "similarity")
B	("lithology" OR "lithologic" OR "oil well" OR "lithostratigraphic")
C	("sequence alignment" OR "sequence comparison")
D	(LIMIT-TO (SUBJAREA , "COMP"))

Tabela 1 – Palavras-chave buscadas na base SCOPUS

As consultas foram realizadas de forma a acumular cada um dos grupos de termos e critérios de busca. Na Tabela 2 abaixo estão os resultados obtidos para cada consulta realizada.

Grupo de palavras-chave	Quantidade de publicações
A	3340436
A AND **B**	14745
A AND **B** AND **C**	4
A AND **B** AND **C** AND **D**	2

Tabela 2 - Resultados obtidos pelas buscas por palavras-chave na base SCOPUS

Nota-se que apenas duas publicações na área de computação relacionam todos os temas centrais deste trabalho. É nesta observação que se encontra sua relevância, através da proposta de uma nova metodologia para correlação de poços de petróleo através de alinhamento de sequências baseadas em litologia.

1.2 Objetivo

O objetivo principal deste trabalho é propor uma nova metodologia para sugestão de poços de petróleo correlatos a partir de clusterização de dados, utilizando como medida de distância o resultado do alinhamento de sequências geradas a partir da litologia de tais poços.

1.3 Estrutura da dissertação

Este trabalho está organizado da seguinte maneira: o Capítulo 2 faz uma revisão na bibliografia e introduz os conceitos básicos necessários para esta pesquisa. O Capítulo 3 apresenta a motivação da pesquisa, mostra trabalhos relacionados e descreve em detalhes a metodologia proposta. No capítulo 4 são apresentados experimentos e resultados obtidos, além de uma discussão geral e, finalmente, no Capítulo 5 são feitas as considerações finais desta pesquisa, assim como sugestões para trabalhos futuros.

2
Conceitos básicos

Este capítulo apresenta de forma breve alguns conceitos importantes relacionados a esta pesquisa, divididos entre construção de poços de petróleo, alinhamento de sequências de DNA e métodos de clusterização (agrupamento) de dados.

2.1
Construção de poços de petróleo

O petróleo, do latim *petrus* (pedra) e *oleum* (óleo), quando em estado líquido é uma substância oleosa, inflamável, de densidade quase sempre menor que a da água, coloração geralmente entre castanho claro e preto. Possui diversos compostos químicos em sua formação, principalmente os hidrocarbonetos (ROSA, CARVALHO e XAVIER, 2006), que foram gerados através da transformação de matérias orgânicas, depositados em terrenos sedimentares, cobertos por sedimentos, e permanecendo sob ação do tempo, pressão e temperatura específicas (PETROBRAS, 2006).

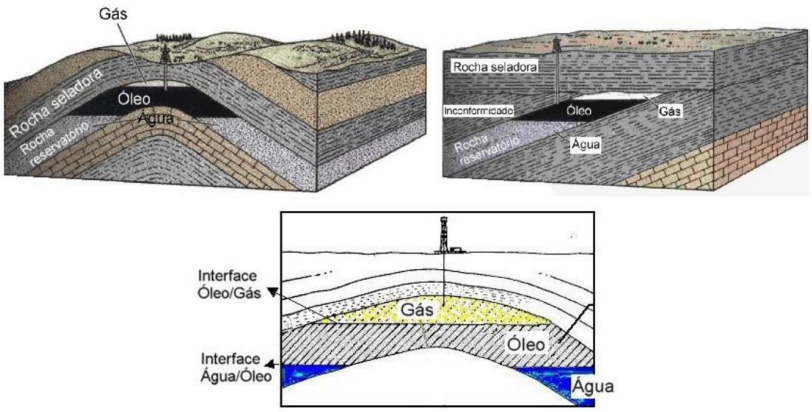


Figura 1 - Sistemas petrolíferos (PETROBRAS, 2006)

Quando essa mistura de compostos químicos contém um percentual maior de moléculas pequenas, nas condições normais de temperatura e pressão, então o estado físico do hidrocarboneto é gasoso (gás natural). Caso contrário, o estado físico é líquido.

Os hidrocarbonetos são gerados nas rochas geradoras, migram para as rochas reservatórios, podendo ser através de rachaduras (falhas tectônicas) e finalmente ficam presos por rochas selantes (através da formação de trapas) como visto na Figura 1. Poço de petróleo é o termo usado para definir uma perfuração na superfície terrestre destinada a produção de hidrocarbonetos (petróleo e/ou gás natural).

A construção de um poço de petróleo é uma tarefa de extrema complexidade, que tem como objetivo preparar um poço de petróleo para posterior desenvolvimento da produção. A construção de um poço de petróleo pode ser dividida basicamente em: perfuração, revestimento/cimentação e completação.

2.1.1 Perfuração

O objetivo desta etapa é atingir o reservatório, onde está localizado o petróleo, através de perfuração utilizando brocas que podem variar entre 5 e 42 polegadas de diâmetro, acoplada à extremidade de uma coluna de tubos, como mostrado na Figura 2.

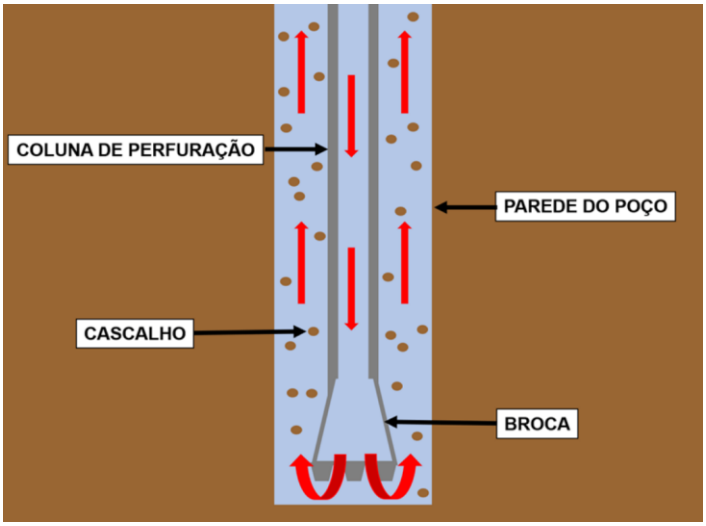


Figura 2 - Operação de perfuração (PELIANO e NEGRÃO, 2018)

2.1.2
Revestimento e cimentação

Após a perfuração de um intervalo, são realizados o revestimento e a cimentação, de modo a manter uma estrutura rígida e segura, permitindo acesso de equipamentos por toda extensão do poço, desde seu topo até o objetivo final.

O revestimento se dá através da descida de diversos tubos metálicos que se encaixam e, posteriormente, ocorre a cimentação, para fixação destes tubos junto às rochas, como ilustrado na Figura 3.

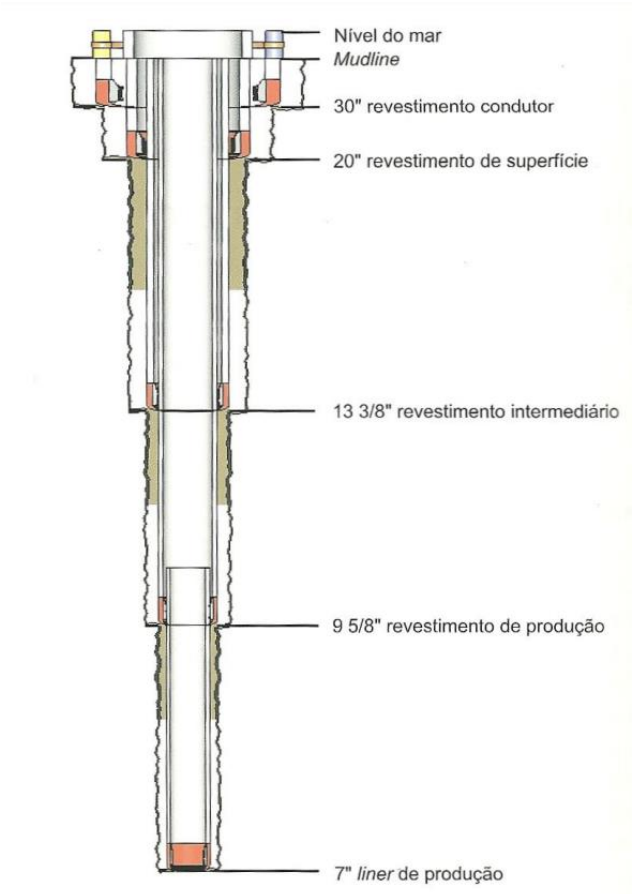


Figura 3 - Revestimentos de um poço típico (ROCHA e DE AZEVEDO, 2019)

A cimentação é um processo sensível aos tipos de rochas encontrados ao longo de todo o intervalo a ser revestido. Dadas as características de porosidade e permeabilidade, resistência a ácidos, coeficiente de expansão, condutividade térmica, entre outras, uma rocha pode demandar volume de pasta de cimentação

maior ou menor que outras, ou ainda demandar determinados aditivos para melhorar a adesão do cimento à parede do poço aberto (OLIVEIRA, GUIMARÃES e MANZOLI, 2016). Desta forma, nesta etapa da construção do poço, o estudo da composição do solo é determinante não apenas para o planejamento das operações, como também para a seleção de materiais e substâncias químicas na formulação das pastas de cimento.

2.1.3 Completação

Nesta etapa o poço é habilitado para produzir óleo ou gás de maneira segura e econômica. Durante a completação equipamentos são instalados na cabeça de poço e pequenos orifícios são criados na região do tubo de revestimento que passa pela zona de produção, de modo a permitir que o poço produza o fluido desejado, que pode ser óleo ou gás (MARTINS, HAMACHER e ACCIOLY, 2014).

2.1.4 Evolução tecnológica

Ao longo do último século houve muitas evoluções tecnológicas na área de construção de poços de petróleo, principalmente os poços *offshore*. Enquanto, no início do século XIX, o desafio era construir poços em águas ultra-rasas (até 30 metros), hoje o desafio é construir poços de petróleo em águas ultra-profundas, com reservatórios com mais de 1500 metros de profundidade (OLIVEIRA, GUIMARÃES e MANZOLI, 2016).

No Brasil, a Petrobras iniciou suas atividades *offshore* nas décadas de 1960-1970. Desde então, muitas pesquisas foram elaboradas de modo a tornar o processo de construção de um poço de petróleo cada vez mais rápido, seguro e barato.

Com a descoberta dos campos do Pré-sal em 2006, ao longo da costa sudeste do Brasil, novos desafios foram lançados e mais pesquisas foram iniciadas com o objetivo de produzir petróleo em locais de acesso limitado (300 quilômetros da costa), com lâminas d'água superiores a dois mil metros e reservatórios com mais de 5 mil metros abaixo do leito marinho, incluindo uma camada de sal com

aproximadamente 2 mil metros de espessura, como pode ser visto na Figura 4 (PETROBRAS, 2021).

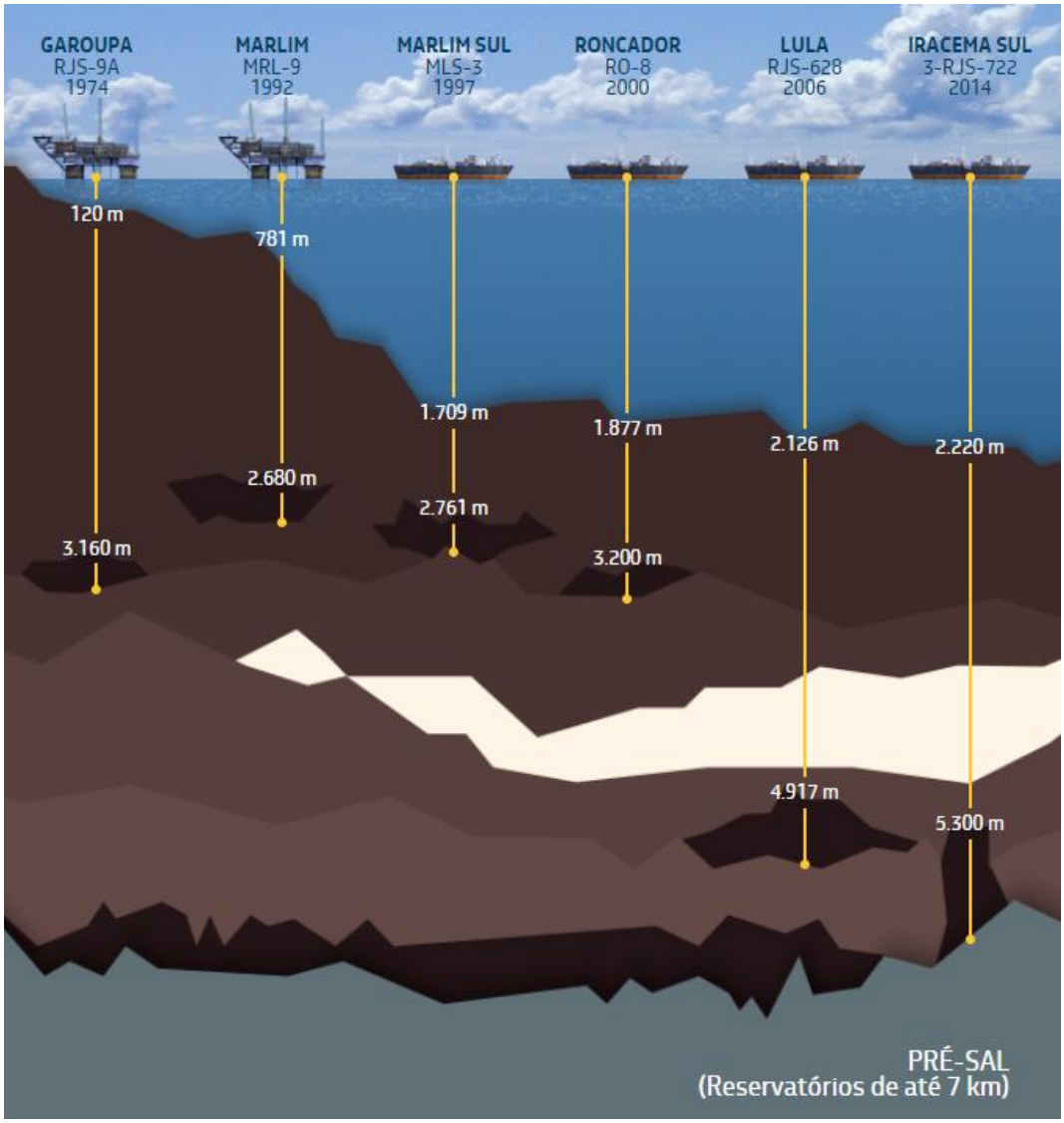


Figura 4 - Evolução tecnológica de construção de poços de petróleo em águas ultra profundas (PETROBRAS, 2021)

Todas essas evoluções demandaram a criação de novos materiais e ferramentas (PERETA, 2015), capazes de realizar as atividades de construção de poços. Alguns recursos possuem custo elevado e são de difícil aquisição, seja por contratações longas ou pelo próprio prazo de fornecimento (*lead-time*) (NEVES, LAZZARINI, *et al.*, 2012). Desta forma, se faz necessária cada vez mais uma previsibilidade sobre quais materiais e equipamentos serão necessários para construção de poços em um horizonte de dois ou mais anos.

2.2
Litologia

As camadas abaixo da superfície são formadas por diversos compostos sólidos, formados por um ou mais minerais. Esses compostos recebem o nome de rocha e a seu estudo é dado o nome de litologia. Este estudo tem por finalidade identificar a composição, estrutura e propriedades físicas da rocha. A partir dele é possível identificar os reservatórios de petróleo (LEITE e GATTASS, 2012). Um outro significado para o termo litologia é a própria identificação do tipo de rocha. E é com este o sentido que este trabalho fará referências ao termo.

O perfil litológico de um poço apresenta todas as rochas encontradas ao longo de sua extensão, sendo cada trecho identificado pelo tipo de rocha, medida de topo e de base. Na Figura 5 podemos ver um exemplo de um perfil litológico de um poço de petróleo, com seus diversos trechos e, para cada um, a definição do tipo de rocha e as medidas de topo e base.

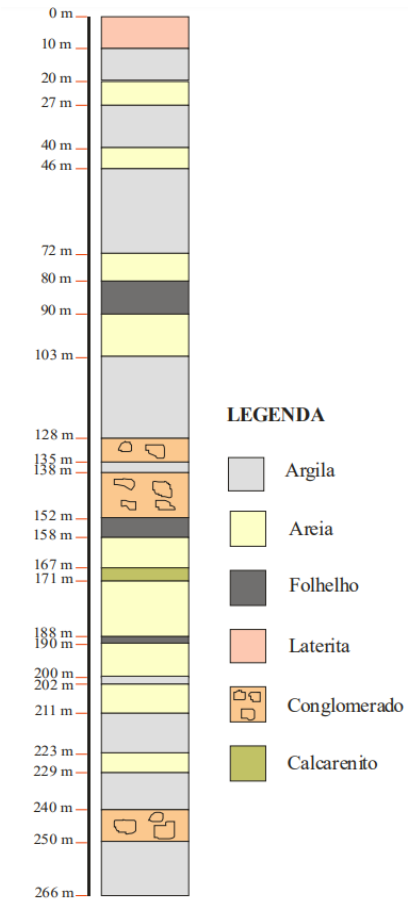


Figura 5 - Perfil litológico de um poço de petróleo (MATTA, ALMEIDA, *et al.*, 2004)

O perfil litológico pode ser obtido durante a perfuração de um poço, geralmente de duas formas (THOMAS, 2004):

- Análise de perfis elétricos: através de sensores diversas características são coletadas das rochas. Essas características são disponibilizadas em gráficos para que geólogos façam uma análise manual a fim de identificar os tipos de rocha e, assim, elaborar o perfil litológico do poço.
- Testemunhagem: é um processo de coleta de amostras das rochas ao longo do poço que está sendo perfurado. Estas amostras são coletadas em cilindros, chamados testemunhos. Por se tratarem de amostras das próprias rochas, a identificação da litologia se faz de maneira direta. A desvantagem desta técnica é que ela é muito cara e, por isso, é executada apenas em poços específicos. Na Figura 6 podemos observar alguns testemunhos coletados de poços de petróleo.

Figura 6 - Testemunhos de poços de petróleo (acervo pessoal)

Já para os poços ainda não perfurados, principalmente em campos de exploração, o perfil litológico pode ser obtido através da interpretação sísmica da região em que o poço será perfurado (JARVIS e SAUSSUS, 2009). Esta interpretação se dá através de uma análise estratigráfica, que fornece uma sequência de camadas de rochas que foram depositadas umas sobre as outras ao longo do tempo (PINHEIRO, 2014). Na Figura 7 pode ser observado um exemplo de interpretação sísmica.

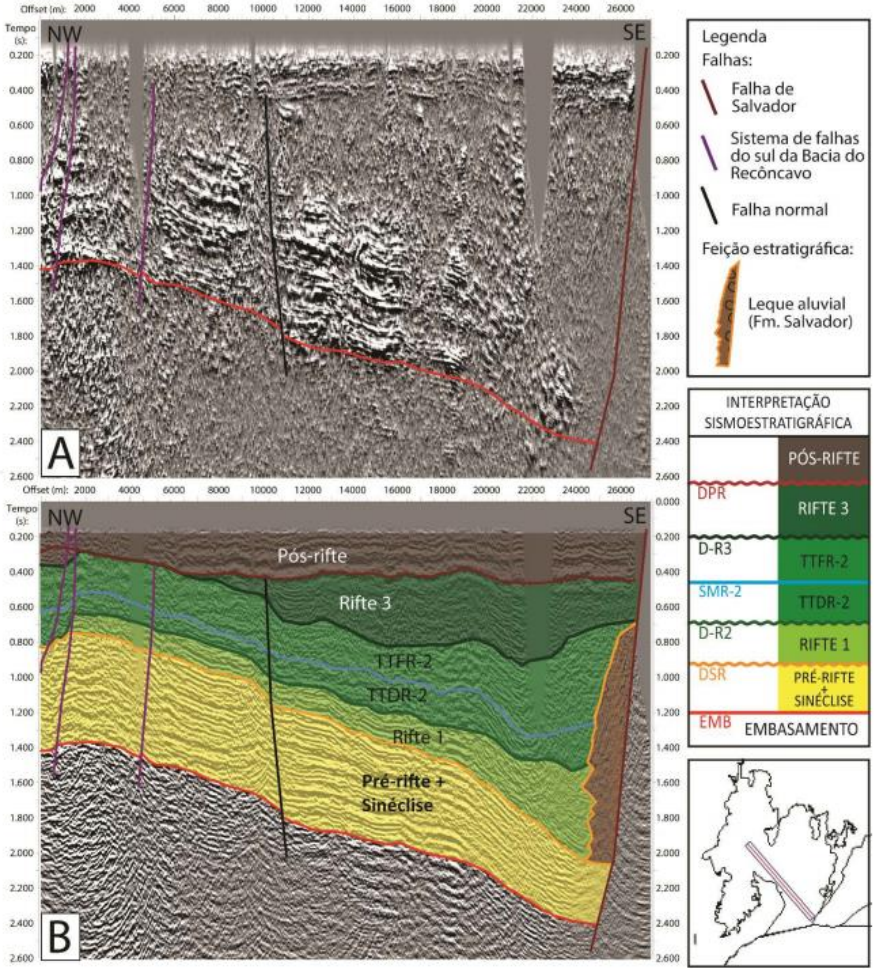


Figura 7 - Interpretação sísmica (VILAS BOAS, SOUZA e HOLZ, 2018)

A partir da interpretação sísmica alinhada ao estudo de poços de correlação e dados secundários, geólogos geram o quadro de previsão geológica (QPG), que indica os prováveis intervalos de formações geológicas do poço a ser perfurado (MATTOS, 2015). Essas informações se assemelham muito ao perfil litológico de um poço já perfurado. Um exemplo de um QPG pode ser visto na Figura 8.

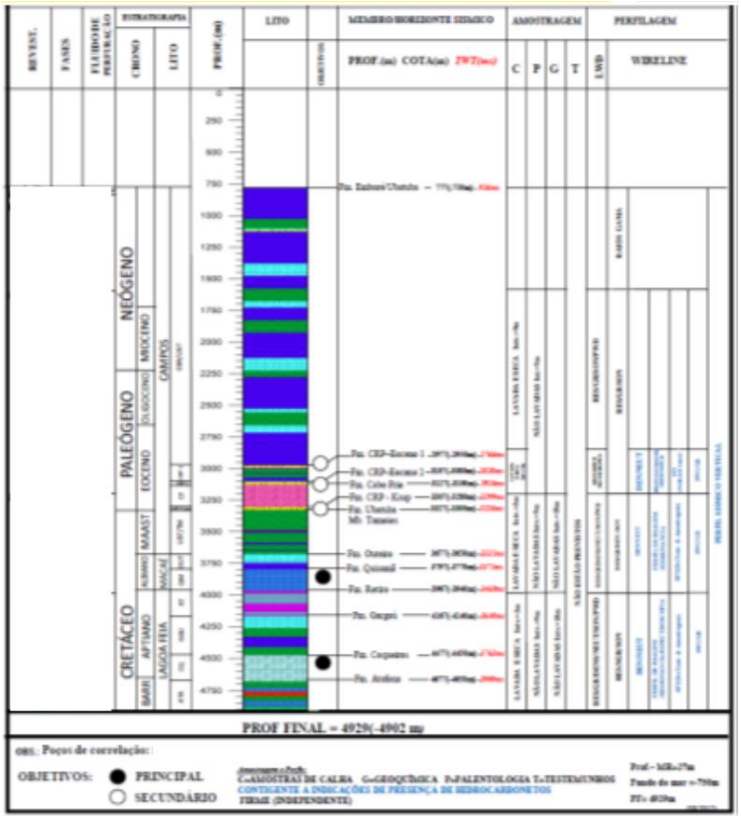


Figura 8 - Quadro de previsão geológica (MASCULO, 2012)

2.2.1
Correlação de poços de petróleo

A correlação de poços de petróleo é uma tarefa de extrema importância e é geralmente realizada com a junção de diversos elementos disponibilizados aos geólogos, como perfis elétricos e estruturais, que devem ser realizados em poços sem revestimento, pois os sensores necessitam ter contato com as paredes do poço, e perfis radioativos, que podem ser realizados em poços já revestidos (SOUZA, 2014). Existem diversos perfis geofísicos: Perfil Raios Gama (GR), Perfil Potencial Espontâneo (SP), Perfil Sônico (DT), Perfil Resistividade (ILD), Perfil Porosidade Neutrônica (NPHI), Perfil Resistência (RE) e Perfil Densidade (RHOB). Outros elementos também são utilizados por geólogos na tarefa de correlação de poços, como as propriedades geológicas e petrofísicas (argilosidade, porosidade, permeabilidade e saturação), além de dados sísmicos.

Com os perfis geofísicos de dois ou mais poços alinhados aos seus perfis litológicos e tendo suas escalas uniformizadas, geólogos podem realizar, de forma manual, a interpretação e, consequentemente, a correlação entre estes poços. Na

Figura 9 abaixo podemos observar uma correlação entre dois poços utilizando os perfis geofísicos Raios Gama (GR) em vermelho, Potencial Espontâneo (SP) em azul e Resistência (RE) em verde, e perfis litológicos.

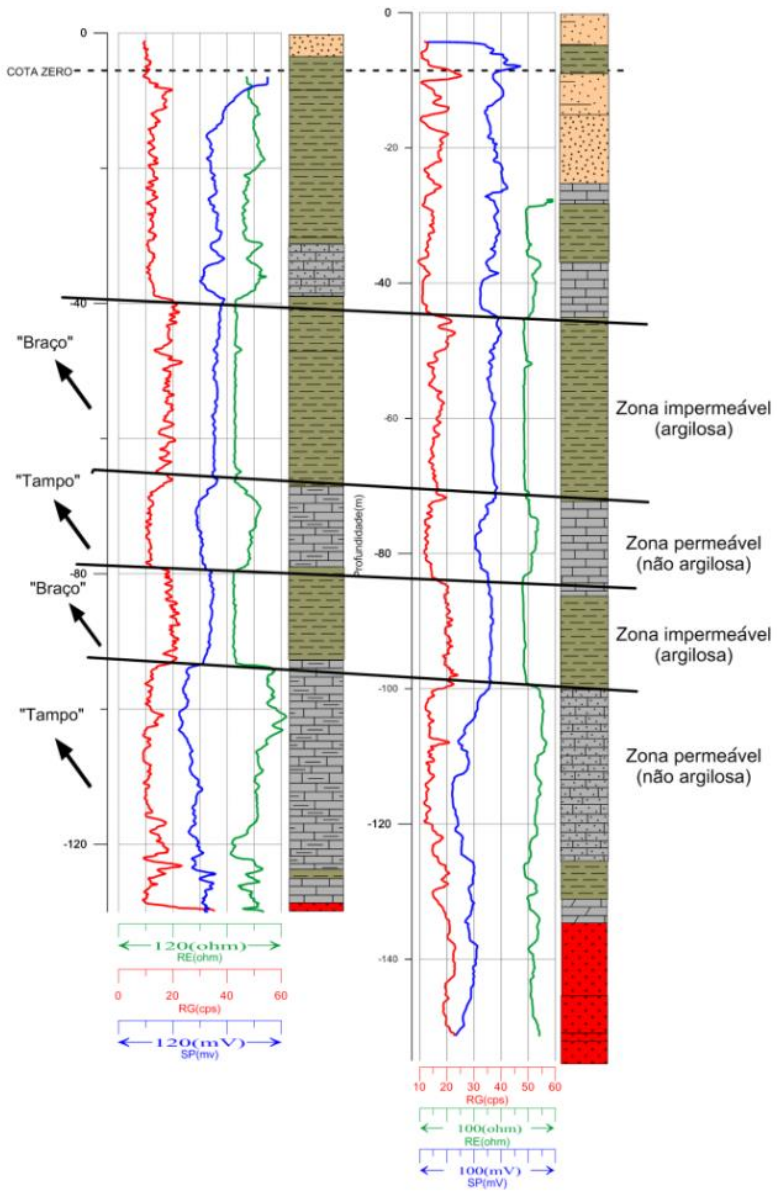


Figura 9 - Correlação de poços utilizando perfis geofísicos e perfis litológicos (FREIMANN, ALVES e SILVA, 2014)

2.3 Alinhamento de sequências de DNA

O DNA, ou ácido desoxirribonucleico, é uma molécula que codifica informações genéticas e é responsável por coordenar o funcionamento e desenvolvimento dos seres vivos (DE SOUZA, 2014). Ela é representada por uma

sequência de letras que representam os nucleotídeos que a compõem: A (Adenina), C (Citosina), T (Timina) e G (Guanina) (DARNELL, LODISH e BALTIMORE, 1990). Um exemplo de uma sequência de DNA pode ser visto na Figura 10.

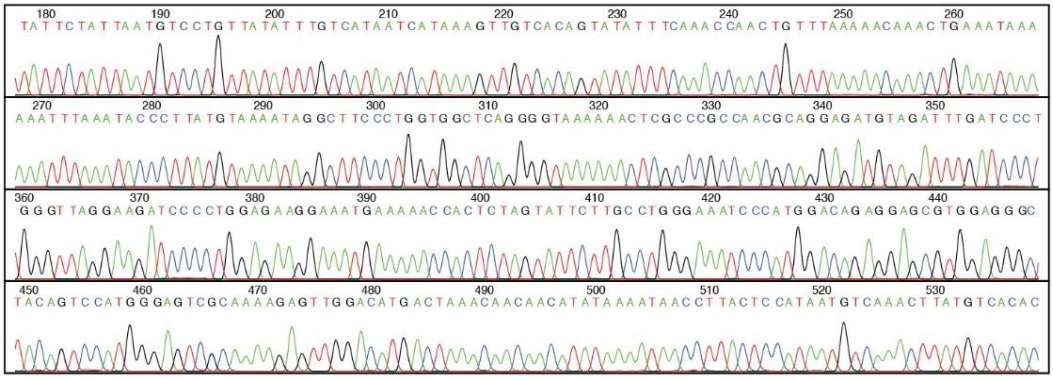


Figura 10 - Sequência de DNA (GREEN, 2021)

A Bioinformática é uma área da computação responsável por empregar técnicas e ferramentas da computação para resolução de problemas de biologia, como o alinhamento de sequências de DNA (BALDI e BRUNAK, 1998), que é o processo de comparação de duas ou mais sequências de nucleotídeos, representadas como sequência de caracteres, com o intuito de verificar suas similaridades, através da identificação de pontos em comum e pontos de divergência entre as sequências (IOSTE, 2016).

De acordo com (SETUBAL e MEIDANIS, 1997) um alinhamento α entre duas sequências s e t é um par de sequências (s' e t') obtidas a partir de s e t, respectivamente, através da inserção de espaços ($-$). O alinhamento final deve satisfazer as seguintes condições:

1. $|s'| = |t'|$, ou seja, as sequências s' e t' possuem o mesmo comprimento.
2. A remoção de todos os espaços de s' retorna s.
3. A remoção de todos os espaços de t' retorna t.
4. Se $s'[i] = -$ então $t'[i] \neq -, \forall i$, e vice versa.

A Figura 11 abaixo ilustra o alinhamento entre duas sequências $s =$ TGGTCT e $t =$ ATCGTT.

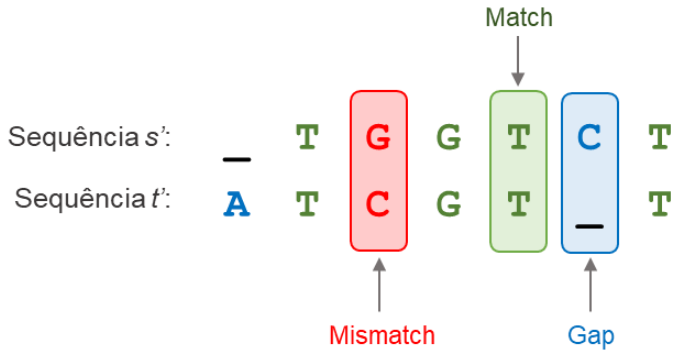


Figura 11 - Alinhamento de sequências com *gap* e *mismatch*

Ainda de acordo com (SETUBAL e MEIDANIS, 1997) a similaridade é o maior *score* entre todos os possíveis alinhamentos entre duas sequências. Um sistema de pontuação é composto por um par (p, g) cujos membros podem ser definidos como a função $p: \Sigma \times \Sigma \mapsto \mathbb{R}$, usada para atribuir pontuações (pesos) para cada par de caracteres alinhados, e uma penalidade para espaços g, sendo geralmente definido como um valor negativo. Com este sistema de pontuação é possível atribuir um valor numérico (*score*) para cada alinhamento possível. Adicionamos $p(a, b)$ cada vez que há um pareamento correto (*match*) entre a e b em α. E adicionamos g toda vez que o caractere a é pareado com um espaço $(-)$. A soma total é o *score* de α, chamado de $score(\alpha)$. Baseando-se neste sistema de pontuação, a similaridade entre duas sequências s e t pode ser descrita como na equação (1) abaixo:

$$sim(s, t) = \max_{\alpha \in A(s,t)} score(\alpha), \qquad (1)$$

onde $A(s, t)$ é o conjunto de todos os alinhamentos possíveis entre s e t.

O algoritmo proposto neste trabalho se baseia no algoritmo Needleman-Wunsch (NEEDLEMAN e WUNSCH, 1970). Esse algoritmo faz o alinhamento global entre duas sequências, através da construção de uma matriz de pontuações (*scores*), que podem ter valores positivos (quando há pareamentos) ou negativos (quando há diferenças ou gaps entre os elementos). A Figura 12 abaixo mostra um exemplo de matriz de pontuação preenchida para as sequências da Figura 11, além do resultado do algoritmo, destacado em amarelo.

	t	T	G	G	T	C
s	0	-1	-2	-3	-4	-5
A	-1	-1	-2	-3	-4	-5
T	-2	0	-1	-2	-2	-3
C	-3	-1	-1	-2	-3	-1
G	-4	-2	0	0	-1	-2
T	-5	-3	-1	-1	1	0

Figura 12 - Matriz resultado do algoritmo Needleman-Wunsch para as sequências $s = ATCGT$ e $t = TGGTC$ com pontuações de $match = 1$, $mismatch = -1$ e $gap = -1$

Na década seguinte foi proposto o algoritmo Smith-Waterman para solucionar o problema do alinhamento local entre duas sequências. Este algoritmo é uma adaptação do algoritmo de Needleman-Wunsch, logo, também é um algoritmo de programação dinâmica. A principal diferença entre os dois é que na matriz de pontuações os valores que porventura fossem negativos são transformados em zero, o que permite que alinhamentos locais possam ser encontrados (SMITH e WATERMAN, 1981).

	t	T	G	G	T	C
s	0	0	0	0	0	0
A	0	0	0	0	0	0
T	0	1	0	0	1	0
C	0	0	0	0	0	2
G	0	0	1	1	0	1
T	0	1	0	0	2	1

Figura 13 - Matriz resultado do algoritmo Smith-Waterman para as sequências $s = ATCGT$ e $t = TGGTC$ com pontuações de $match = 1$, $mismatch = -1$ e $gap = -1$

Os algoritmos de programação dinâmica se propõem a encontrar o melhor alinhamento possível (pode existir mais de um alinhamento ótimo). Uma limitação destes algoritmos é que, pelo fato deles terem performance de pior caso na ordem de $O(m * n)$, sendo m e n os tamanhos das duas sequências, podem se tornar lentos dependendo dos tamanhos das sequências, além de não serem muito escaláveis quando aplicados ao problema de alinhamento de múltiplas sequências. Diversas adaptações foram realizadas nesses algoritmos de forma a obter um resultado

biologicamente mais significativo, ter um melhor aproveitamento de espaço ou diminuir a complexidade computacional (SETUBAL e MEIDANIS, 1997). Outras técnicas também foram propostas para atender essas necessidades, como métodos baseados em heurística e utilizando bancos de dados de sequências (palavras ou k-tuplas) já catalogadas, que não atingem uma solução ótima, mas sim aproximada.

2.3.1 Algoritmo Needleman-Wunsch

O algoritmo Needleman-Wunsch foi escolhido para o contexto deste trabalho por se tratar de um algoritmo exato para alinhamento global entre duas sequências, com resultado ótimo e tendo performances de espaço e desempenho adequadas para o contexto não biológico a que este trabalho se propõe (sequências relativamente pequenas).

Por se tratar de um algoritmo de programação dinâmica (DP, de *dynamic programming*), ele segue a mesma definição para se modelar um problema a ser resolvido por DP, que consiste em quatro passos (CORMEN, LEISERSON e RIVEST, 1989):

1. Caracterizar a estrutura de uma solução ótima;
2. Recursivamente definir o valor de uma solução ótima;
3. Calcular o valor de uma solução ótima, normalmente de baixo para cima (*bottom-up*); e
4. Construir a solução ótima por meio das informações calculadas.

Para o alinhamento entre as sequências s e t de tamanhos m e n, respectivamente, o algoritmo Needleman-Wunsch considera que se um pareamento entre dois caracteres s_i e t_j, das sequências s e t, faz parte de uma solução ótima para o subproblema de alinhamento entre as subsequências $s_1 \dots s_i$ e $t_1 \dots t_j$, então o pareamento entre s_i e t_j faz parte de uma solução ótima para o alinhamento global entre as sequências s e t. Após esta caracterização é feita a criação da matriz bidimensional, semelhante à construída para o método *Dot Plot*, onde os valores serão calculados e armazenados. Para isto, primeiramente são definidos os parâmetros de pontuação $score(a, b)$ e gap penalty (g). A função $score(a, b)$ determina as pontuações específicas para alinhamento entre quaisquer dois

elementos possíveis do alfabeto a ser considerado no alinhamento entre s e t, tanto para quando houver pareamento (*match*, quando $a = b$), quanto para quando não houver (*mismatch*, quando $a \neq b$). Após esta definição, será montada uma matriz F de dimensões $m + 1 \times n + 1$, da seguinte maneira:

$$\begin{cases} F_{i,0} = g * i, \forall i \\ F_{0,j} = g * j, \forall j \end{cases} \quad (2)$$

Para as sequências $s =$ TGGTC e $t =$ ATCGT a matriz F, gerada a partir da equação (2), pode ser visualizada na Figura 14, abaixo.

	t	T	G	G	T	C
s	0	-1	-2	-3	-4	-5
A	-1					
T	-2					
C	-3					
G	-4					
T	-5					

Figura 14 - Matriz inicial para as sequências s = ATCGT e t = TGGTC com gap penalty = -1

O preenchimento da matriz já inicializada F se dá da seguinte forma, iniciando em $i = j = 1$:

$$F_{i,j} = Max \begin{cases} F_{i-1,j-1} + score(s_i, t_j) \\ F_{i-1,j} + g \\ F_{i,j-1} + g \end{cases}, para\ i \in [1, m]\ e\ j \in [1, n] \quad (3)$$

Pela equação (3), para cada célula é definida uma pontuação máxima através de cálculos envolvendo os valores das células adjacentes que já foram preenchidas anteriormente (acima, abaixo e na diagonal cima-esquerda), como pode ser visto na Figura 15 abaixo.

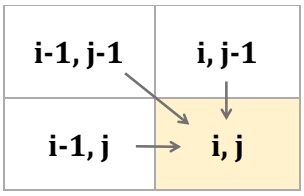


Figura 15 - Células utilizadas para o cálculo do valor de uma célula (i, j) na matriz do algoritmo Needleman-Wunsch

Após toda matriz F ser preenchida, temos a pontuação (score) final e o alinhamento global ótimo resultante para as sequências s e t. Para isto, deve ser percorrido o melhor caminho a partir da célula $F_{m,n}$ e ir voltando (*traceback*) determinando se serão inseridos espaços (*gaps*) ou não de acordo com a forma que a pontuação daquela célula foi gerada.

	t	T	G	G	T	C
s	0	-1	-2	-3	-4	-5
A	-1	-1	-2	-3	-4	-5
T	-2	0	-1	-2	-2	-3
C	-3	-1	-1	-2	-3	-1
G	-4	-2	0	0	-1	-2
T	-5	-3	-1	-1	1	0

Figura 16 - Matriz resultado do algoritmo Needleman-Wunsch contendo traceback para as sequências s = ATCGT e t = TGGTC com pontuações de match=1, mismatch=-1 e gap=-1

Na Figura 16 acima pode ser verificado que o alinhamento ótimo para as sequências $s =$ TGGTC e $t =$ ATCGT, tendo como parâmetros $match = 1$, $mismatch = -1$ e $gap\ penalty = -1$, é definido como $s' =$ _TGGTC e $t' =$ ATCGT_.

2.4 Clusterização de dados

O problema de agrupamento de dados (*Data Clustering*), aqui chamado de clusterização de dados, é um assunto amplamente estudado dentro dos campos de

mineração de dados (*Data Mining*) e aprendizado de máquinas (*Machine Learning*), por conta das inúmeras aplicações em sumarizações, aprendizado, segmentação e direcionamento de mercado (KUMAR, 2014).

A clusterização de dados é um conjunto de técnicas de aprendizado de máquina não-supervisionado (*Unsupervised Learning*), que é um tipo de abordagem cuja finalidade é identificar padrões em um conjunto de dados (*dataset*) não rotulados. O objetivo da clusterização de dados é agrupar os dados de maneira que objetos de um mesmo grupo (ou *cluster*) sejam muito similares entre si, enquanto objetos de *clusters* distintos sejam muito diferentes. A Figura 17 mostra o resultado de uma clusterização de um *dataset* com duas dimensões (variáveis) e cinco *clusters* encontrados. Diferentemente do problema de classificação, onde as classes são previamente conhecidas, no problema de clusterização de dados essas classes também precisam ser descobertas (GAN, MA e WU, 2007).

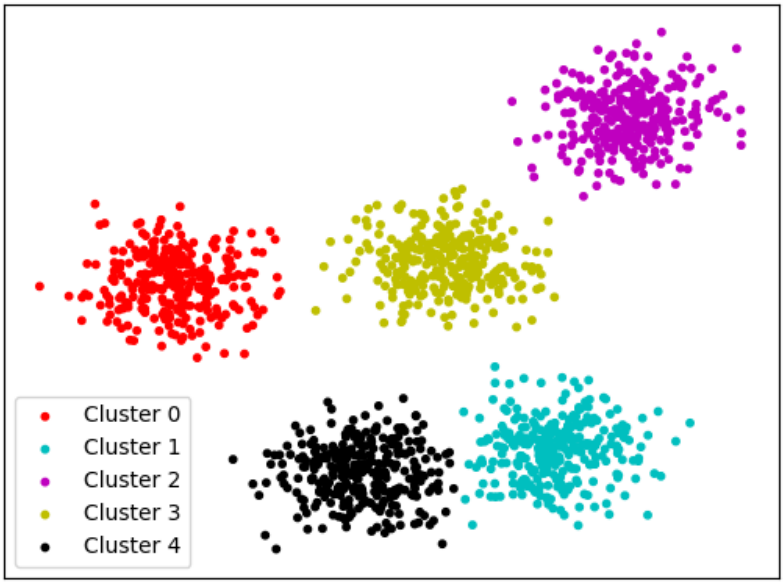
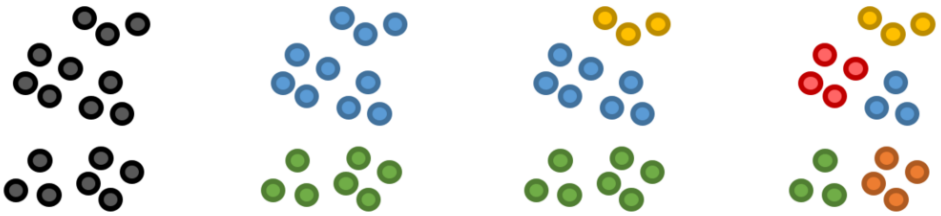


Figura 17 - Exemplo de clusterização de dados num espaço bidimensional (ISSARANE, 2021)

Segundo Hair Jr., Black, Babin & Anderson (2015) a clusterização em geral envolve três etapas. A primeira é a definição de uma medida de similaridade entre os objetos do *dataset* de forma a permitir a determinação de quantos grupos existem na amostra. A segunda etapa é a do processo de clusterização em si, onde os objetos são particionados em grupos (*clusters*). Este processo pode ser notado na Figura 18

abaixo, onde inicialmente (a) os objetos não possuem nenhuma categorização (*cluster*) e, à medida que o processo é executado, os objetos vão sendo alocados em *clusters* (b, c e d). O passo final é traçar o perfil das variáveis para determinar sua composição.



(a) (b) (c) (d)

Figura 18 – Alocação de objetos em *clusters* com k = 5

Como dito anteriormente, a medida de similaridade (ou distância) entre os objetos do *dataset* desempenha um importante papel no processo de clusterização de dados. Medidas de similaridade, coeficientes de similaridade, medidas de dissimilaridade ou distâncias são usadas para descrever quantitativamente a similaridade ou dissimilaridade entre dois objetos ou dois *clusters* (GAN, MA e WU, 2007). Em geral, distância e similaridade são conceitos antagônicos, ou seja, quanto maior a similaridade entre dois objetos, menor a distância entre eles. Existem diversas formas de se medir a distância entre dois objetos, sendo a distância euclidiana a mais utilizada. A equação (4) abaixo mostra como calcular a distância euclidiana entre dois pontos $x = (x_1, x_2, \dots, x_n)$ e $y = (y_1, y_2, \dots, y_n)$ em um espaço de n dimensões:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (4)$$

Hruschka e Ebecken (2003) definiram formalmente o problema de clusterização. Considerando um conjunto de n objetos $X = \{X_1, X_2, \dots, X_n\}$, onde cada $X_i \in \mathbb{R}^\rho$ é um vetor de ρ medidas reais. Os objetos precisam ser agrupados em

k *clusters* disjuntos $C = \{C_1, C_2, \ldots, C_k\}$, onde k é o número de *clusters*. As seguintes condições precisam ser respeitadas:

$$C_1 \cup C_2 \cup \ldots \cup C_k = X; \quad (5)$$

$$C_i \neq \emptyset, \forall i, 1 \leq i \leq k; \quad (6)$$

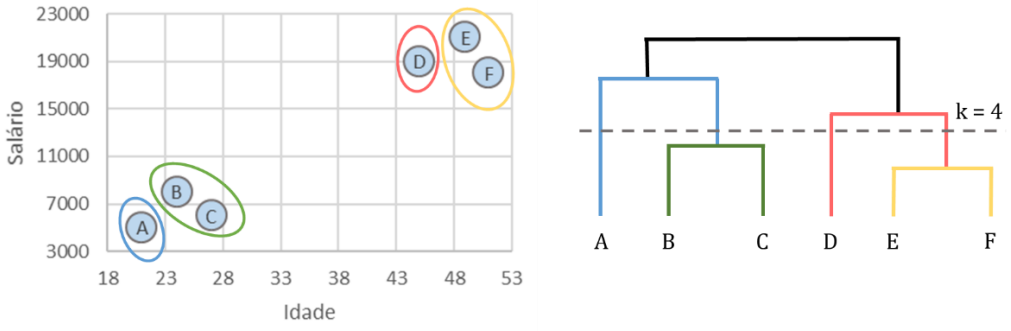
$$C_i \cap C_j = \emptyset, \forall i \neq j, 1 \leq i \leq k, 1 \leq j \leq k. \quad (7)$$

Pela equação (5) temos que todos os objetos do *dataset* inicial estão espalhados entre todos os *clusters*. Já a equação (6) mostra que todos os *clusters* devem ter pelo menos um objeto e, por fim, a equação (7) condiciona que nenhum objeto participa de mais de um *cluster*, o que significa que são *clusters* disjuntos.

A categorização dos algoritmos de clusterização é uma tarefa complexa, pois muitas características se sobrepõem, colocando alguns algoritmos em mais de uma categoria (BERKHIN, 2006). Mesmo assim, alguns autores (HAN, KAMBER e PEI, 2012) classificam os algoritmos nessas principais grandes categorias: métodos hierárquicos, métodos baseados em partição (particionais), métodos baseados em densidade e métodos baseados em grade, além de outros.

2.4.1 Métodos Hierárquicos

Segundo (HAN, KAMBER e PEI, 2012) os métodos hierárquicos criam uma decomposição hierárquica do conjunto de objetos a serem agrupados, e podem ser classificados como aglomerativos ou divisivos. Os aglomerativos utilizam uma abordagem *bottom-up*, considerando cada objeto como sendo um *cluster* ($k = n$). Após isto, junções são realizadas a cada passo do algoritmo, até que reste apenas um *cluster* no final. Geralmente é determinado um k para que o algoritmo pare quando houver k *clusters*. De forma análoga, os divisivos utilizam uma abordagem *top-down*, onde começam considerando que todos os objetos fazem parte do mesmo cluster e então vão, a cada passo do algoritmo, quebrando um *cluster* existente em *clusters* menores. Da mesma forma, geralmente um valor para k é definido de modo a indicar o critério de parada do algoritmo .



(a) (b)

Figura 19 - Clusterização hierárquica em um *dataset* de 6 registros e $k = 4$

Um grande problema de algoritmos hierárquicos é que uma vez que um passo tenha sido executado, a operação não pode ser desfeita. Ou seja, se um objeto foi definido como sendo de um *cluster* em um determinado ramo da hierarquia, ele não poderá ser transferido para um *cluster* contido em outro ramo. A Figura 19 acima mostra (a) um exemplo de clusterização realizada por método hierárquico para um *dataset* formado por 6 objetos e $k = 4$, através de (b) um dendrograma.

2.4.2 Métodos Particionais

Dado um *dataset* de n objetos, um método de partição constrói k partições destes dados, onde cada partição representa um *cluster* e $k \leq n$. Ou seja, ele divide os dados em k grupos contendo, cada um, ao menos um objeto e um elemento central, chamado de centroide. A maioria dos métodos particionais é baseada em distância. Diferentemente dos métodos hierárquicos, os algoritmos que implementam os métodos particionais utilizam uma técnica de realocações iterativas, que tentam melhorar o particionamento movendo objetos de um *cluster* para outro, caso este possua distância menor para um *cluster* vizinho, ou seja, para o centroide de um *cluster* vizinho (HAN, KAMBER e PEI, 2012).

O critério geral de um bom particionamento é que os objetos no mesmo *cluster* são muito similares, enquanto os objetos em diferentes *clusters* são muito diferentes. Geralmente é computacionalmente proibitivo atingir uma otimização global em um método de clusterização por partições, pois seria exigido uma enumeração exaustiva de todas as partições possíveis. No entanto, a maioria dos

algoritmos que implementam esta técnica utilizam métodos heurísticos comuns, como abordagens gulosas, como os algoritmos K-means e K-medoids, que melhoram progressivamente a qualidade do agrupamento. Essas heurísticas funcionam bem para encontrar *clusters* planos esféricos em *datasets* de tamanho pequeno a médio. Para encontrar *clusters* com formas complexas ou em *datasets* muito grandes, os métodos de particionamento precisam ser estendidos (HAN, KAMBER e PEI, 2012), como no caso dos métodos de clusterização baseados em dimensões, como o DBSCAN, por exemplo. A Figura 20 abaixo mostra exemplos de formas que algoritmos particionais não conseguiriam descobrir e como métodos baseados em dimensões se comportam.

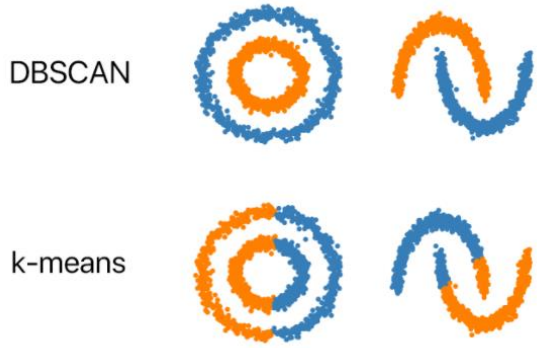


Figura 20 - Comportamento de algoritmos de clusterização baseados em densidade e partições para formas não planas circulares (adaptado de (MATTT, 2020))

2.4.2.1 K-means

K-means, proposto inicialmente por (MACQUEEN, 1967), é o algoritmo de clusterização particional mais simples e mais utilizado (KUMAR, 2014). Por sua simplicidade, facilidade de implementação e eficiência empírica, continua sendo, mesmo após mais de 50 anos de sua publicação, um dos algoritmos mais utilizados para clusterização de dados (JAIN, 2009). Ele começa através da escolha de *k* pontos representativos como sendo os centroides iniciais de cada um dos *k clusters*. Cada objeto do *dataset* é, então, atribuído a um dos *clusters* de acordo com uma medida particular de similaridade escolhida. Com os *clusters* formados, os centroides de cada *cluster* são atualizados. O algoritmo então repete estes últimos

dois passos até que os centroides não tenham uma atualização significativa ou que algum critério de convergência seja satisfeito (KUMAR, 2014). O K-means é um algoritmo guloso que garantidamente converge para um mínimo local, mas a minimização de sua função de pontuação é conhecida como NP-Difícil (MAHAJAN, NIMBHORKAR e VARADARAJAN, 2012).

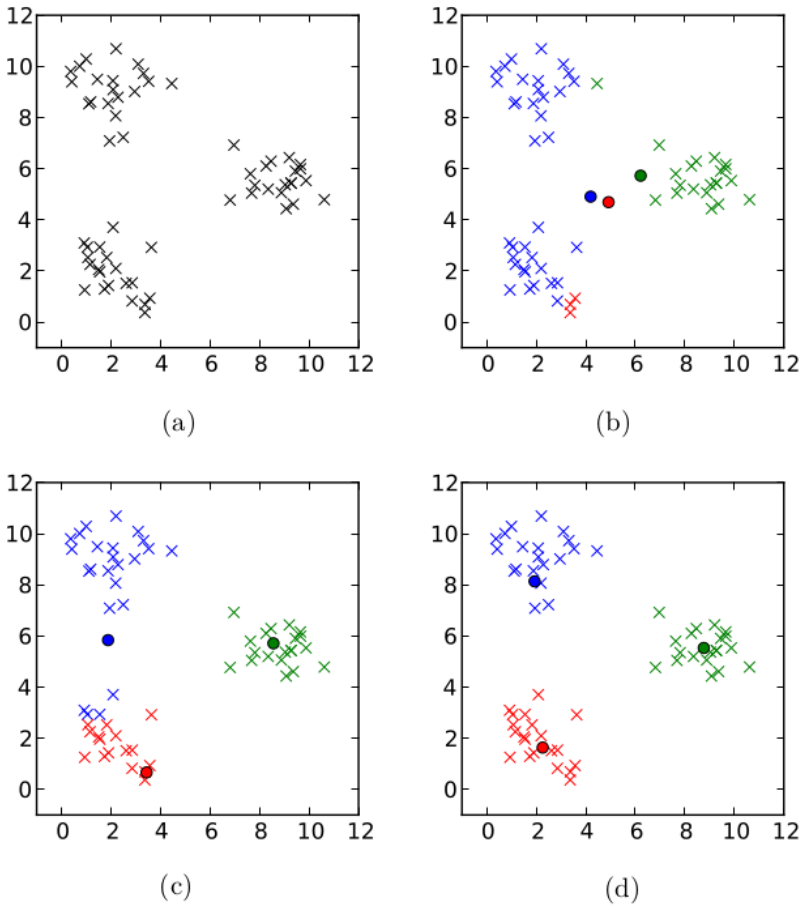


Figura 21 - Convergência do algoritmo K-means em um *dataset* bidimensional com k = 3 (adaptado de (ARNOLD, 2021))

Na Figura 21 acima pode ser visto (a) um *dataset* bidimensional representado por todos os seus objetos (denotados por ×), depois (b) o passo inicial do algoritmo K-means, com a definição dos centroides (denotados por ○) e, finalmente, em (c) e (d) podem ser visualizados resultados de duas iterações do algoritmo, com a atualização dos centroides a cada iteração e a realocação dos objetos aos *clusters*, representados por cores iguais às dos centroides. O algoritmo K-means pode ser descrito como na Figura 22 abaixo (KUMAR, 2014).

Algoritmo 1: Clusterização K-means	
Entrada:	
	k: número de *clusters* a serem calculados
	D: *dataset* contendo n objetos
Saída: um conjunto de n *clusters*	
Método:	
1:	Defina k novos pontos para serem os centroides dos *clusters*
2:	**Repita**:
3:	(Re)atribua cada um dos n objetos ao *cluster* que aquele objeto seja mais parecido, baseado na média dos objetos presentes naquele cluster
4:	Recalcule o centroide de cada *cluster*, através do recálculo da média de objetos cada cluster
5:	**Até** que o critério de convergência seja satisfeito

Figura 22 - Clusterização K-means (adaptado de (HAN, KAMBER e PEI, 2012))

Segundo (HAN, KAMBER e PEI, 2012) uma técnica baseada em centroides utiliza o centroide de um cluster C_i para representa-lo. Conceitualmente, o centroide de um *cluster* é o seu ponto central. O centroide pode ser definido de várias maneiras, como pela média entre os objetos ou pela mediana entre os objetos (*medoids*). A diferença entre um objeto $p \in C_i$ e c_i, que é a representação do *cluster*, é medida por $dist(p, c_i)$, onde $dist(x, y)$ é uma medida de distância, que pode ser, de diferentes formas, como a distância Euclidiana, distância de Manhattan, similaridade por cossenos ou outras (KUMAR, 2014). Em geral, algoritmos K-means utilizam a distância Euclidiana. Dependendo do valor de k e da medida de similaridade escolhidos, diferentes resultados podem ser obtidos através da clusterização utilizando K-means. A função objetivo empregada pelo K-means é chamada de Soma dos Erros Quadrados (*Sum of Squared Errors*, ou SSE) ou Soma Residual de Quadrados (*Residual Sum of Squares*, ou RSS) e na equação (8) abaixo pode ser vista sua formulação matemática para obtenção da soma dos erros quadrados de um *cluster* C qualquer.

$$SSE(C) = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2 \quad (8)$$

Um fator importante que impacta a eficácia do algoritmo K-means é a definição do valor k (KUMAR, 2014). Uma seleção aleatória dos valores de k geralmente resulta em agrupamentos de baixa qualidade. Por este motivo, alguns métodos foram propostos para definição de um melhor valor de k. Um método simples é simplesmente definir $k = \sqrt{\frac{n}{2}}$, resultando em uma média de $\sqrt{2n}$ objetos por *cluster* (HAN, KAMBER e PEI, 2012). Outro método é o chamado de "cotovelo" (*elbow method*), que se baseia no entendimento de que o aumento da quantidade de *clusters* pode ajudar a reduzir a soma da variância dos objetos contidos em cada cluster. Isto é porque ter mais *clusters* permite obter grupos com objetos que são mais similares entre si. No entanto, o efeito de redução da soma das variâncias dos objetos internos de cada *cluster* pode diminuir caso muitos *clusters* sejam formados, pois a divisão de um *cluster* coeso em dois resulta apenas em uma pequena redução.

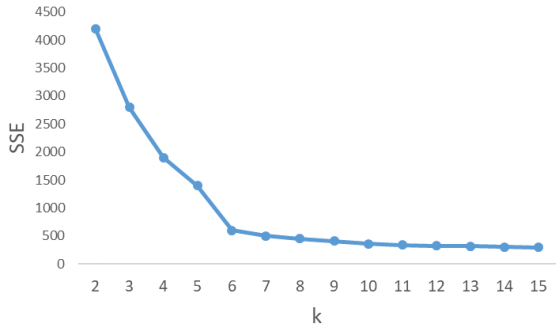


Figura 23 - Método "cotovelo" (*elbow*) para obtenção do valor de k

O método de "cotovelo" utiliza a Soma dos Erros Quadrados (SSE) entre os objetos e seus respectivos centroides para cada valor de $k \in [2, n]$, geralmente traçados em um gráfico de linha tendo k no eixo x. Após isto, seleciona-se o valor k na posição do gráfico em que a linha traçada forma um "cotovelo", ou seja, o ponto onde a partir dele não há ganho significativo em relação ao SSE. A Figura 23 acima mostra um exemplo de gráfico traçado e o ponto de formação do "cotovelo",

para obtenção do valor de k, que no exemplo é $k = 6$. A Figura 23 acima mostra um exemplo de gráfico traçado e o ponto de formação do "cotovelo", para obtenção do valor de k, que no exemplo é $k = 6$.

2.4.2.2 K-medoids

Segundo (HAN, KAMBER e PEI, 2012) o algoritmo K-means é sensível a objetos que estão muito distantes da maioria dos dados (*outliers*) pois quando estes *outliers* são atribuídos a um *cluster*, podem distorcer radicalmente o valor médio do *cluster*. Isso afeta a atribuição de outros objetos aos *clusters*. Este efeito é intensificado devido ao uso da função de erro quadrático da equação (8). O método K-medoids foi proposto por (KAUFMAN e ROUSSEEUW, 1987) como uma modificação ao algoritmo K-means com objetivo de diminuir sua sensibilidade em relação aos *outliers*. No método K-medoids, em vez de utilizar o valor médio dos objetos em um *cluster* como ponto de referência, através da criação de novos objetos, deve-se escolher objetos reais para representar os *clusters*, tendo um objeto representativo para cada *cluster*. Cada objeto restante é então atribuído ao *cluster* do qual o objeto representativo é o mais semelhante. O método de particionamento é então executado com base no princípio de minimização da soma das dissimilaridades entre cada objeto $\boldsymbol{p}$ e seu objeto representativo correspondente. Ou seja, um critério de erro absoluto é usado, definido como na equação (9), abaixo:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{o_i}) \tag{9}$$

Onde E é a soma do erro absoluto para todos os objetos $\boldsymbol{p}$ no dataset, e $\boldsymbol{o_i}$ é o objeto representativo de C_i. Esta é a base do algoritmo K-medoids, que agrupa n objetos em k *clusters* através da minimização do erro absoluto. Para $k > 1$, o algoritmo k-medoids é um problema NP-Difícil (HAN, KAMBER e PEI, 2012).

O algoritmo PAM (*Partitioning Around Medoids*) é uma implementação do método K-medoids muito popular (HAN, KAMBER e PEI, 2012). Ele aborda o problema de forma gulosa e iterativa. Assim como no algoritmo K-means, os objetos representativos (aqui chamados de *medoids*) iniciais $\{o_1, o_2, \ldots, o_k\}$ são

escolhidos aleatoriamente, mas no caso do K-medoids tais objetos devem necessariamente estar contidos no *dataset*. Então, é verificado se a qualidade do agrupamento melhora ao substituir um *medoid* ($\boldsymbol{o_j}, 1 \leq j \leq k$) por um objeto não representativo qualquer ($\boldsymbol{o_{random}}$). Todas as substituições possíveis são testadas, calculando a distância de cada objeto $\boldsymbol{p}$ ao objeto mais próximo do conjunto $\{o_1, \dots, o_{j-1}, o_{random}, o_{j+1}, \dots, o_k\}$ e utiliza-se esta distância para atualizar a função de custo. O processo iterativo de substituição de objetos representativos por outros objetos continua até que a qualidade do *cluster* resultante não possa ser melhorada por nenhuma substituição. Essa qualidade é medida por uma função de custo da dissimilaridade média entre um objeto e o objeto representativo de seu *cluster* (HAN, KAMBER e PEI, 2012). A formalização do algoritmo PAM pode ser vista na Figura 24 abaixo.

Algoritmo 2: Clusterização K-medoids (PAM)

Entrada:

k: número de *clusters* a serem calculados

D: *dataset* contendo n objetos

Saída: um conjunto de n *clusters*

Método:

1: Escolha arbitrariamente k objetos de D para serem os objetos representativos iniciais dos *clusters*

2: **Repita**:

3: Atribua cada um dos $(n - k)$ objetos remanescentes de D ao *cluster* que aquele objeto seja mais parecido

4: Randomicamente escolha um objeto não representativo, $\boldsymbol{o_{random}}$.

5: Calcule o custo total, S, de trocar o objeto representativo $\boldsymbol{o_j}$ por $\boldsymbol{o_{random}}$.

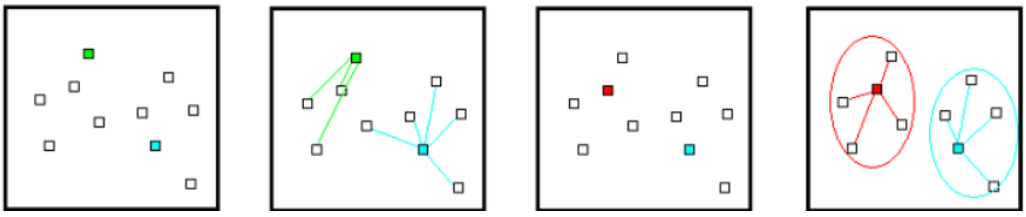
6: **Se** $S < 0$ **então** troque $\boldsymbol{o_j}$ por $\boldsymbol{o_{random}}$ para formar o novo conjunto de k objetos representativos.

7: **Até** que o critério de convergência seja satisfeito

Figura 24 - Clusterização K-medoids (PAM) (adaptado de (HAN, KAMBER e PEI, 2012))

Cada vez que ocorre uma reatribuição, uma diferença no erro absoluto, E, é adicionada à função de custo. Portanto, a função de custo calcula a diferença no

valor do erro absoluto se um objeto representativo atual for substituído por um objeto não representativo. O custo total da troca é a soma dos custos incorridos por todos os objetos representativos. Se o custo total for negativo, então o_j é substituído ou trocado por o_{random} porque o erro absoluto real E é reduzido. Se o custo total for positivo, o objeto representativo atual, o_j, é considerado aceitável e nada é alterado na iteração. A Figura 25 abaixo mostra os passos de uma iteração do algoritmo PAM, com (a) a escolha inicial de objetos representativos, (b) a atribuição de todos os outros objetos aos *clusters* mais parecidos, (c) a escolha de um objeto não representativo de forma aleatória, representado pela cor vermelha, e (d) a troca do *medoid* para o objeto escolhido no passo anterior.



(a) (b) (c) (d)

Figura 25 - Passos do algoritmo K-medoids (adaptado de (NAVER, 2009))

3
Metodologia

Este capítulo apresenta uma análise preliminar do problema que motivou a realização deste trabalho, e então o detalhamento da metodologia proposta nesta pesquisa.

3.1
Análise preliminar do problema

Uma grande companhia de exploração e produção de petróleo, como a Petrobras®, necessita de muito planejamento para a correta contratação de materiais e equipamentos necessários à construção de poços de petróleo. Este planejamento está inserido em um cenário de muitas incertezas em relação a como serão os projetos dos poços a serem construídos nos próximos anos. Esta incerteza se dá pois grande parte dos poços a serem construídos estão em áreas exploratórias, onde existe pouca informação sobre os reservatórios.

Se este planejamento não for realizado corretamente, as contratações de materiais e equipamentos serão superestimadas, o que acarretará em aumento do estoque e, consequentemente, aumento do custo de construção do poço de petróleo, que no final se traduz em prejuízo financeiro para a companhia.

Por este motivo, geólogos realizam correlações entre poços de modo obter uma estimativa para o projeto de construção de um poço futuro. Desta forma, as áreas responsáveis pela aquisição de materiais e equipamentos conseguem iniciar os processos de contratações, que frequentemente levam mais de um ano para serem realizadas e depois levam mais de um ano para que o material ou equipamento seja efetivamente entregue.

Por ser uma quantidade muito elevada de poços (mais de 6400 no cenário pesquisado), a correlação visual realizada por geólogos de forma manual se torna muito dependente da experiência da pessoa que está realizando, além de não ser possível correlacionar um novo poço com uma grande quantidade de outros poços,

permitindo, assim, uma melhor estimativa de projeto de construção do novo poço. Diante disso, há a necessidade de criação de ferramentas que auxiliem geólogos na tarefa de sugestão de poços correlatos para servirem de modelos para um novo poço.

Este trabalho tem como objetivo propor uma metodologia capaz de oferecer automaticamente sugestões de poços correlatos a um dado poço com base em clusterização baseada no alinhamento de sequências, que são geradas a partir dos dados litológicos de tais poços.

3.2
Trabalhos Relacionados

Como visto na introdução deste trabalho, existem poucos outros trabalhos que relacionam o assunto de correlação de poços de petróleo baseada em sua litologia com o assunto de alinhamento de sequências. Nas buscas realizadas na base SCOPUS o primeiro trabalho encontrado que relaciona esses assuntos é o intitulado *Ontologies applied to lithologic correlation problem within the petroleum geology domain* (Ontologias Aplicadas ao Problema de Correlação Litológica no Domínio da Geologia do Petróleo), de (GARCIA, CARBONERA e ABEL, 2013). Neste trabalho os autores investigaram abordagens para o problema da correlação litológica, no domínio da Estratigrafia Sedimentar. Primeiramente, segundo os autores, geólogos descrevem visualmente *corpos de rocha*, tal como o *testemunho de sondagem*, através da identificação das *fácies sedimentares*, por meio do apontamento de todos os atributos visuais nelas presentes, como tamanho do grão, cor, litologia, arredondamento, esfericidade, dentre outras.

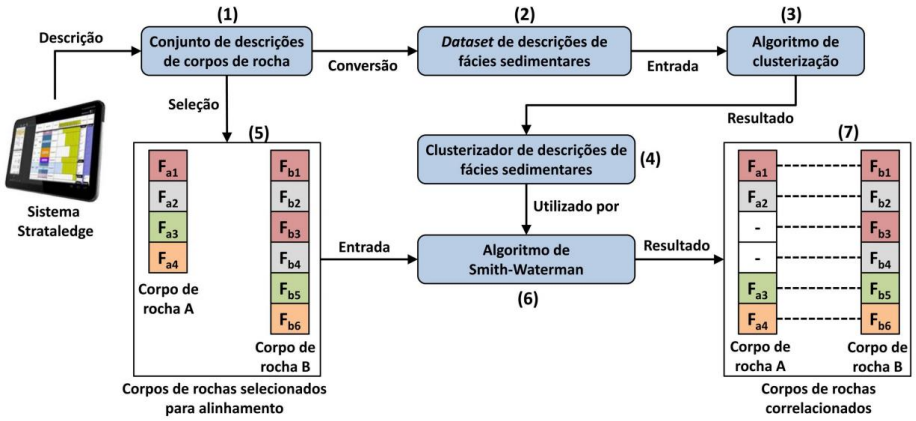


Figura 26 - Representação dos procedimentos realizados na abordagem proposta por Garcia, Carbonera e Abel (GARCIA, CARBONERA e ABEL, 2013)

O trabalho mencionado acima (GARCIA, CARBONERA e ABEL, 2013) tem como objetivo realizar a correlação poço a poço utilizando alinhamento local de sequências, ou seja, sem a geração de um *score* final como resultado do alinhamento entre as sequências. Além disso, as sequências alinhadas são geradas a partir dos identificadores dos *clusters* obtidos após a clusterização das fácies sedimentares, através de seus atributos. A Figura 26 acima mostra o esquema proposto no estudo mencionado.

Algumas dificuldades foram encontradas no trabalho acima, como o processo utilizado para a geração das sequências, através de um método de aprendizado de máquina não supervisionado. Isso faz com que a representação de cada poço em uma sequência dependa de todo o *dataset* e não apenas dos dados do próprio poço, além de depender de uma correta parametrização por parte de especialistas para que cada atributo possua um peso apropriado a ser utilizado pelo algoritmo de clusterização. Finalmente, o grupo de trabalho concluiu que a clusterização para obtenção de sequências não foi uma abordagem apropriada, pois técnicas de agrupamento de dados geram um julgamento binário (as fácies sedimentares são ou não são similares entre si e, portanto, devem ou não ser identificadas pelo mesmo *cluster*). Outra dificuldade encontrada foi a escassez de dados para realização de experimentos, tendo menos de 10 poços disponíveis para experimentação.

Para que a tarefa de clusterização de poços através das sequências geradas a partir da litologia dos poços possa ser realizada, é necessário que exista uma forma de calcular a distância entre dois poços. Por este motivo, o alinhamento local entre sequências não se mostra uma abordagem apropriada, pois não é gerado um *score* final, mas sim um conjunto de *scores* máximos, cada um iniciando uma subsequência alinhada. Além disso, é importante que um determinado poço possa ter sua sequência representativa sendo gerada a partir apenas de seus próprios dados, e não dependendo de todo o *dataset*. Estes fatores motivaram a criação de uma nova metodologia para correlação automática de poços através de alinhamento de sequências baseadas em litologia. Esta metodologia é explicada a seguir.

Outros trabalhos relacionados ao tema foram escritos pelo mesmo grupo de estudo do trabalho apresentado acima e são muito semelhantes entre si, não necessitando de uma explicação adicional.

3.3
Metodologia Proposta

A sugestão de poços correlatos através do alinhamento de sequências baseadas em litologia, proposta neste trabalho, não consiste apenas de um único método, mas sim de uma metodologia composta por seis etapas. O diagrama apresentado na Figura 27 abaixo mostra o sequenciamento destas etapas, que são tratadas em detalhes nas seções seguintes.

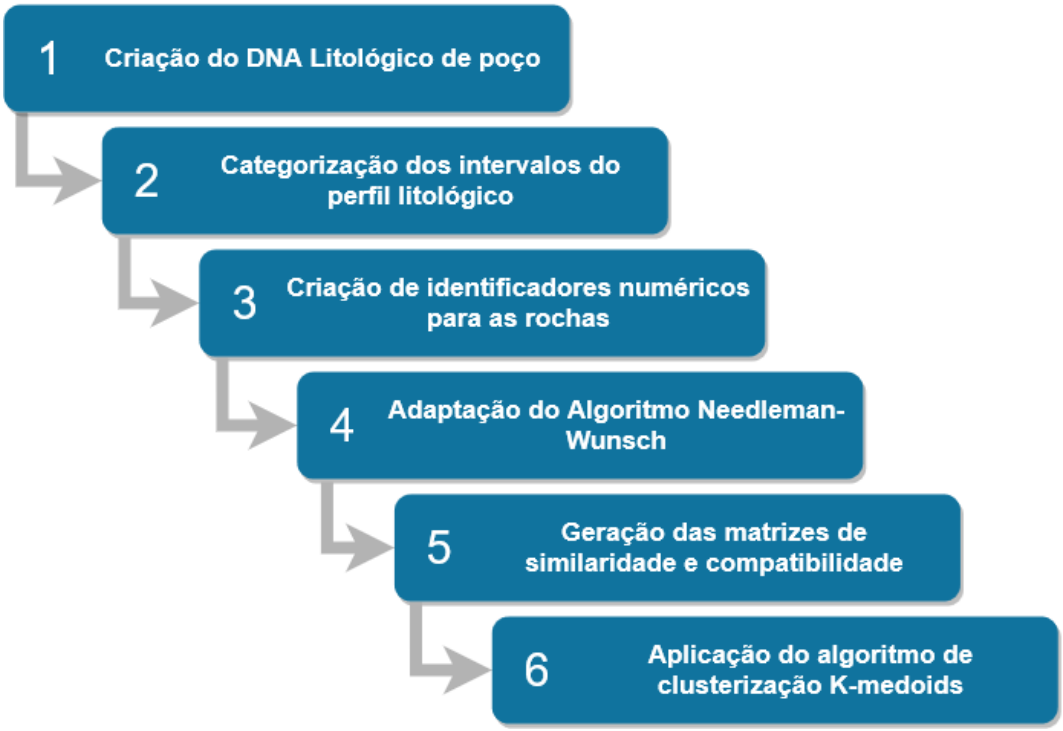


Figura 27 - Metodologia proposta

3.3.1
Criação do DNA Litológico de poço

Quando o perfil litológico de um poço de petróleo, ou o quadro de previsão geológica, é analisado, uma sequência de intervalos de rochas pode ser identificada. Assim como em uma sequência de DNA, que possui subsequências de letras de um alfabeto próprio, $\{A, C, T, G\}$, para representar os nucleotídeos, é possível transformar o perfil litológico de um poço em um conjunto de subsequências de caracteres. Cada tipo de rocha seria representado por um caractere diferente e o

tamanho da subsequência (quantidade de caracteres iguais contíguos) seria proporcional ao tamanho do intervalo de rocha a ser representado.

Figura 28 - Exemplo de DNA Litológico com base na regra de um caractere para cada metro de intervalo de rocha representado

O conjunto de subsequências resultante desta conversão é chamado, neste trabalho, de **DNA Litológico** do poço. Se para cada metro do perfil litológico do poço um caractere for adicionado ao DNA Litológico deste poço, seu comprimento será equivalente à quantidade de metros de extensão do poço.

Topo (m)	Base (m)	Tipo de rocha
0	3	ARENITO
3	11	FOLHELHO
11	84	ANIDRITA
84	320	ARGILA
320	940	DIAMICTITO
940	2000	ANIDRITA

Tabela 3 - Exemplo de definição de perfil litológico

Tipo de Rocha	Caractere representativo
ARENITO	*a*
FOLHELHO	*f*
ARGILA	*g*
DIAMICTITO	*d*
ANIDRITA	*i*

Tabela 4 - Exemplo de caracteres identificadores de rochas

A Figura 28 acima mostra um exemplo de criação do **DNA Litológico** do poço mostrado na Tabela 3 acima, com base na regra de imprimir um caractere para cada metro de intervalo de rocha a ser representado, sendo cada tipo de rocha identificado por um caractere distinto, como descrito na Tabela 4 acima.

3.3.2 Categorização dos intervalos do perfil litológico

Ainda que o comprimento do **DNA Litológico** resultante seja muito pequeno (poucos milhares de caracteres) quando comparado ao comprimento de sequências de DNA de seres vivos, com bilhões de caracteres em todo o genoma (NATIONAL HUMAN GENOME RESEARCH INSTITUTE, 2021), quando as matrizes de similaridade e compatibilidade forem geradas para todos os poços, o algoritmo que implementa a metodologia proposta neste trabalho poderia se tornar inviável computacionalmente.

Além disso, quando geólogos realizam a tarefa de correlação visual entre dois perfis litológicos de poços, os intervalos não são comparados metro a metro, como a metodologia proposta até aqui sugere, o que geraria muitos *mismatches* em um alinhamento de sequências muito parecidas, porém com intervalos de rochas com comprimentos diferentes.

Tamanho *t* do intervalo (em metros)	Quantidade de caracteres
$t < 5$	1
$5 \leq t < 10$	2
$10 \leq t < 100$	3
$100 \leq t < 500$	4
$500 \leq t < 1000$	5
$t \geq 1000$	6

Tabela 5 - Categorias dos intervalos de rochas para geração de DNA Litológico

Por esta razão, foi realizada uma categorização dos intervalos de um perfil litológico de acordo com o tamanho de cada um, tal que para cada categoria um número específico de caracteres seja atribuído ao DNA Litológico do Poço. As categorias estão descritas na Tabela 5 acima.

Desta forma, o **DNA Litológico** do poço possuirá um comprimento mais próximo da quantidade de intervalos presentes no perfil litológico do poço, o que

tornará o processo de alinhamento de sequências muito mais eficiente e condizente com o que geólogos realizam, na prática. A Figura 29 abaixo mostra o **DNA Litológico** do poço especificado na Tabela 3 com base na categorização de intervalos.

```
In [5]: dna_poco_intervalos_categorizados
Out[5]: 'aggfffggggddddiiiiii'
```

Figura 29 - Exemplo de DNA Litológico do Poço com base na regra de representação de categorias de intervalos representados

3.3.3 Criação de identificadores numéricos para as rochas

Ao analisar os perfis litológicos dos poços do *dataset*, foi verificada a existência de 56 diferentes tipos de rochas. Sendo assim, a utilização caracteres para representar cada tipo de rocha se mostrou inadequada, pois mesmo que fossem definidos caracteres diferentes para cada um dos tipos de rocha, esta abordagem não seria facilmente customizável para outros domínios ou até para um novo *dataset* com mais tipos de rocha existentes, já que necessitaria de intervenção manual para escolher um caractere específico para cada novo tipo de rocha.

Tipo de Rocha	Identificador numérico
ARGILA	4
ARENITO	12
ANIDRITA	28
DIAMICTITO	31
FOLHELHO	55

Tabela 6 - Exemplo de identificadores numéricos para tipos de rocha

Foi introduzido, então, um passo na metodologia para criar uma identificação numérica para cada tipo de rocha e depois o algoritmo Needleman-Wunsch foi reescrito de forma que o alinhamento de sequências seja feito entre dois vetores de números inteiros e não mais entre duas sequências de caracteres.

A Tabela 6 acima mostra exemplos de identificadores para os tipos de rocha encontrados no perfil litológico do poço de exemplo mostrado na Tabela 3. Já a

Figura 30 abaixo mostra um exemplo do **DNA Litológico**, do mesmo poço, gerado com base na categorização de intervalos e dos identificadores de rochas exemplificados na Tabela 6.

```
In [3]: dna_poco_intervalos_categorizados_com_identificadores_rochas
Out[3]: [12, 4, 4, 55, 55, 55, 4, 4, 4, 4, 31, 31, 31, 31, 31, 28, 28, 28, 28, 28, 28]
```

Figura 30 - Exemplo de DNA Litológico do Poço com base na categorização de intervalos utilizando identificadores numéricos para os tipos de rocha

3.3.4
Adaptação do Algoritmo Needleman-Wunsch

O algoritmo Needleman-Wunsch foi adaptado para permitir a obtenção de dois fatores de relação entre poços: similaridade e compatibilidade, que são abordados nos tópicos seguintes, assim como as adaptações realizadas.

3.3.4.1
Similaridade entre poços

O conceito de similaridade entre se baseia no próprio *score* final obtido através do alinhamento entre dois **DNAs litológicos**. Primeiramente foram considerados os valores da Tabela 7 abaixo para os parâmetros de entrada do algoritmo.

Parâmetro	Valor
Match	2
Mismatch	-2
Gap penalty	-1

Tabela 7 - Valores dos parâmetros de entrada para o Algoritmo Needleman-Wunsch adaptado

Desta forma, o *score* máximo que pode ser obtido do alinhamento entre duas sequências s e t ocorre quando elas são exatamente iguais, ou seja, $\mathrm{s} = \mathrm{t}$. Se o comprimento destas sequências for m, então o *score* final será $2m$. Analogamente, o *score* mínimo resultante do alinhamento entre duas sequências ocorre quando

ambas possuem o mesmo comprimento, mas são completamente diferentes, ou seja, $s[i] \neq t[i], \forall i$. Neste caso, o *score* final será $-2m$.

Sendo assim, o *score* final do alinhamento entre duas sequências s e t de comprimentos m e n, respectivamente, com os parâmetros de entrada acima estipulados, ficará sempre no intervalo $[-2(m + n)..2(m + n)]$. A partir deste *score* final, denotado por $F[m, n]$, obtém-se o fator de similaridade entre duas sequências de **DNA litológico** s e t, de comprimentos m e n, respectivamente, como um valor no intervalo $[0..1]$. A equação (10) abaixo mostra a formalização matemática do fator de similaridade.

$$similaridade(s, t) = \frac{\left(\frac{F[m, n]}{m + n}\right) + 1}{2} \quad (10)$$

3.3.4.2 Compatibilidade entre poços

Ainda que dois poços, p_s e p_t, sejam muito similares entre si, a existência, em p_s, de um tipo de rocha não existente em p_t, pode fazer com que os materiais e equipamentos utilizados para construção deste (p_t) não sejam adequados para o outro (p_s). Isto ocorre quando há, em p_s, presença de rochas que demandam materiais ou equipamentos diferenciados tecnologicamente (como, por exemplo, brocas de perfuração especiais), e tais rochas não se assemelham às rochas encontradas no poço p_t. Para resolver este problema, este trabalho propõe a criação do fator de compatibilidade, que busca atribuir uma punição para rochas que aparecem em apenas um dos poços e que não são compatíveis com as rochas presentes no outro poço.

Para isso, primeiramente foi criado um parâmetro de entrada extra para a metodologia, que é a matriz de compatibilidades entre os tipos de rochas. Esta matriz triangular de dimensões $r \times r$ possui, para cada par de tipos de rocha, um valor no intervalo $[0..1]$ representando o grau de compatibilidade entre as duas rochas. Na Figura 31 abaixo há um exemplo de matriz de compatibilidade para um conjunto de cinco tipos de rocha $\{A, C, T, G, X\}$.

	A	C	T	G	X
A	1	.95	.92	.89	.01
C		1	.98	.07	.20
T			1	.93	.10
G				1	.39
X					1

Figura 31 - Exemplo de matriz de compatibilidade entre tipos de rochas

O algoritmo Needleman-Wunsch possui um terceiro passo, que é a realização do *traceback*, no qual obtém as sequências alinhadas. No contexto deste trabalho esta etapa foi adaptada para não obter as sequências alinhadas, mas sim os pares em que houve erro (*gaps* e *mismatches*). A Figura 32 abaixo mostra os erros encontrados através do *traceback* adaptado para o alinhamento das sequências s = TGGTC e t = ATCGT.

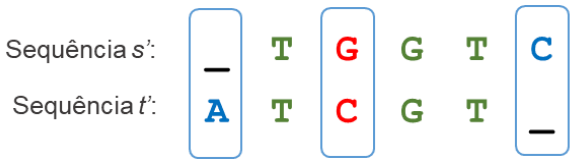


Figura 32 - Exemplo de erros encontrados no alinhamento de duas sequências

A partir desses pares de erros, são formados dois vetores E_s e E_t, contendo, cada um, as rochas que apareceram nos erros referentes a s' e t', respectivamente. Para o exemplo mostrado na Figura 32, tem-se que $E_s = [G, T]$ e $E_t = [A, C]$. Em seguida, novos vetores U_s e U_t são criados, contendo apenas as rochas de E_s e E_t, respectivamente, que não aparecem na outra sequência original, ou seja, em t e s, respectivamente. As equações (11) e (12) abaixo mostram a formalização matemática da construção destes vetores.

$$U_s = \bigcup_{i=1}^{|E_s|} E_s[i] \not\subset t \tag{11}$$

$$U_t = \bigcup_{i=1}^{|E_t|} E_t[i] \not\subset s \tag{12}$$

A quantidade de rochas não exclusivas é somada à quantidade de rochas presentes em *matches* para obter a variável N.

Complementando o exemplo anterior, tem-se que $U_s = \emptyset$ e $U_t = [A]$, pois apenas o caractere A, presente em um erro da sequência t não ocorre na outra sequência, s. Além disso, tem-se que $N = 9$, pois há 3 *matches* (com duas rochas em cada) e 3 rochas não exclusivas nos erros (G, T, C). Em seguida, são criados os vetores W_s e W_t, que armazenam, para toda rocha r presente em U_s e U_t, tuplas contendo, para cada rocha distinta r' presente na outra sequência, a compatibilidade de r em relação à r' e a quantidade de vezes que r' aparece.

Continuando o exemplo anterior, $W_s = \emptyset$, pois $U_s = \emptyset$. Já $W_t = [(0.95, 1), (0.92, 2), (0.89, 2)]$, pois existem, em s, uma ocorrência da rocha C (peso 0.95 em relação à rocha A), duas ocorrências de T (peso 0.92 em relação à rocha A) e duas ocorrências de G (peso 0.89 em relação à rocha A). Os pesos estão descritos na Figura 31 mais acima. Após isto, é criado o vetor único $W = W_s \cup W_t$.

A compatibilidade entre as sequências s e t pode, finalmente, ser calculada de acordo com o parâmetro de entrada "método de compatibilidade", que pode ter três possíveis valores: média ponderada, peso mínimo ou peso máximo.

Média ponderada: nesta opção, a compatibilidade é calculada através da média ponderada entre os pesos e quantidades presentes nas tuplas de W, somado à N e, posteriormente, dividido pela quantidade total de rochas nas duas sequências $(m + n)$. No caso do exemplo anterior, $W = \frac{(\frac{1*0.95 + 2*0.92 + 2*0.89}{1+2+2}) + 9}{10} = 0.9914$.

Peso mínimo: nesta opção, o menor peso obtido é utilizado, independentemente da quantidade. Ele então é multiplicado pela razão entre N e a quantidade total de rochas nas duas sequências $(m + n)$. Ao analisar o exemplo anterior através da compatibilidade por peso mínimo, tem-se que $W = \min(0.95, 0.92, 0.89) \times \frac{9}{10} = 0.801$.

Peso máximo: analogamente à compatibilidade por peso mínimo, a compatibilidade por peso máximo busca o maior peso presente em W, também multiplicado por $n/(m + n)$. Ou seja, $W = \max(0.95, 0.92, 0.89) \times \frac{9}{10} = 0.855$.

O cálculo da compatibilidade pode ser formalizada matematicamente através da equação (13) abaixo, considerando W e método m como entrada.

$$comp(s,t) = \begin{cases} \dfrac{\left(\dfrac{\sum_{i=1}^{|W|} W[i][0] * W[i][1]}{\sum_{i=1}^{|W|} W[i][1]}\right) + N}{m+n}, metodo = "ponderada" \\ \dfrac{min(W[i][0]) \times N}{m+n}, 1 \leq i \leq |W|, metodo = "mínimo" \\ \dfrac{max(W[i][0]) \times N}{m+n}, 1 \leq i \leq |W|, metodo = "máximo" \end{cases} \quad (13)$$

O algoritmo de cálculo de compatibilidade é formalmente descrito na Figura 33 abaixo:

Algoritmo 3: Cálculo de Compatibilidade

Entrada:

s: DNA Litológico do poço p_s de tamanho m

t: DNA Litológico do poço p_t de tamanho n

$metodo$: método de compatibilidade {PONDERADA, MINIMO, MAXIMO}

c: matriz de compatibilidade entre as rochas ($r \times r$)

Saída: fator de compatibilidade entre as sequências s e t

Método:

1: $alinhamento = \boldsymbol{AlgoritmoNeedlemanWunschAdaptado}(s,t)$
2: $pares\_erros = alinhamento.retornaParesComErros()$ $\#traceback$
3: $E_s = E_t = U_s = U_t = W_s = W_t = W = []$
4: $N = 2 \times quantidade\ de\ matches$
5: **para cada** par_{erro} em $\text{pares}_{\text{erros}}$:
6: **se** $\text{par}_{\text{erro}}[0] \neq -$ **então** adiciona $\text{par}_{\text{erro}}[0]$ em E_s
7: **se** $\text{par}_{\text{erro}}[1] \neq -$ **então** adiciona $\text{par}_{\text{erro}}[1]$ em E_t
8: **para cada** rocha em E_s:
9: **se** rocha não existe em t **então** adiciona rocha em U_s **senão** $N = N + 1$
10: **repetir** as linhas **7** e **8** substituindo E_s por E_t, t por s, e U_s por U_t
11: **para cada** $rocha$ em U_s:
12: **para cada** $rocha'$ em $distinct(t)$:
13: compatibilidade = compatibilidades[rocha, $rocha'$]
14: adiciona (compatibilidade, ocorrências de $rocha'$ em t) em W_s

15: **repetir** as linhas **10 à 13** substituindo U_s por U_t, t por s, e W_s por W_t

16: $W = W_s \cup W_t$

17: **se** metodo $=$ PONDERADA **então** retorna $\frac{\left(\frac{\sum_{i=1}^{|W|} W[i][0]*W[i][1]}{\sum_{i=1}^{|W|} W[i][1]}\right)+N}{m+n}$

18: **se** metodo $=$ MÍNIMO **então** retorna $\frac{min(W[i][0]) \times N}{m+n}, 1 \leq i \leq |W|$

19: **se** metodo $=$ MÁXIMO **então** retorna $\frac{max(W[i][0]) \times N}{m+n}, 1 \leq i \leq |W|$

Figura 33 - Algoritmo de cálculo de compatibilidade entre dois DNAs Litológicos

3.3.5
Geração das matrizes de similaridade e compatibilidade

Após a coleta e descaracterização de dados, criação do **DNA Litológico** para todos os poços e adaptações no algoritmo Needleman-Wunsch, é possível construir as matrizes de similaridade e compatibilidade. Primeiramente foram criadas quatro matrizes $n \times n$, sendo n a quantidade de poços: uma para conter o fator de similaridade poço-a-poço e outras três para conter fatores de compatibilidade poço-a-poço para os tipos média ponderada, mínimo e máximo.

Com as matrizes criadas, foi feito o preenchimento de cada uma delas. Para cada par de poços p_s e p_t, foi executado o algoritmo Needleman-Wunsch adaptado e armazenado (tanto na posição $[s, t]$ quanto na posição $[t, s]$) os valores de similaridade, compatibilidade por média ponderada, compatibilidade por mínimo e compatibilidade por máximo, em cada uma das matrizes. Ao final do processo, as quatro matrizes estão completamente preenchidas e disponíveis para a execução da próxima etapa da metodologia.

3.3.6
Aplicação do algoritmo de clusterização K-medoids

Antes da aplicação do método de clusterização K-medoids é necessário identificar a quantidade ideal de *clusters*, denominado por k. Podem ser utilizados diversos métodos para esta finalidade. Neste trabalho foi utilizado o método do cotovelo (*Elbow Method*) explicado anteriormente. Este método foi reimplementado de modo a utilizar a diferença entre 1 (um) e a matriz de similaridades criada no passo anterior, gerando, assim, uma matriz de distâncias

entre os poços. A partir destes dados, é possível, para cada valor possível de $k = [2, \dots, k_{max}]$, aplicar o algoritmo K-medoids e calcular, para cada *cluster*, a soma dos quadrados das distâncias entre os elementos do *cluster* e o seu *medoid*. Ao final esses valores são novamente somados para, então, chegar ao valor da Soma dos Erros Quadrados (*Sum of Squared Errors*, ou SSE).

Com o o valor de *k* encontrado, são escolhidos aleatoriamente *k* poços como os *medoids* iniciais e então aplicado o algoritmo K-medoids passando como parâmetros a matriz de distâncias e os *k medoids* iniciais.

4
Experimentos e resultados

Este capítulo demonstra os resultados obtidos após a implementação da metodologia proposta e sua execução em um *dataset* de 120 poços marítimos.

4.1
Coleta de dados

Para a aplicação da metodologia proposta, foi coletado um *dataset* contendo dados de perfis litológicos de 120 poços de petróleo marítimos já perfurados, obtidos junto à Petrobras®. Os poços estão localizados em 15 blocos de concessão (junto à ANP) diferentes, sendo cada um dos blocos representado por oito poços, exatamente. Os nomes dos blocos, poços e rochas, além das medidas de topo e base de cada intervalo do perfil litológico foram descaracterizados, de modo a garantir a confidencialidade dos dados obtidos. Esses 120 poços estão espalhados geograficamente pela costa sudeste brasileira, como pode ser observado na Figura 34 abaixo, sendo cada um deles colorido de acordo com o bloco de concessão.

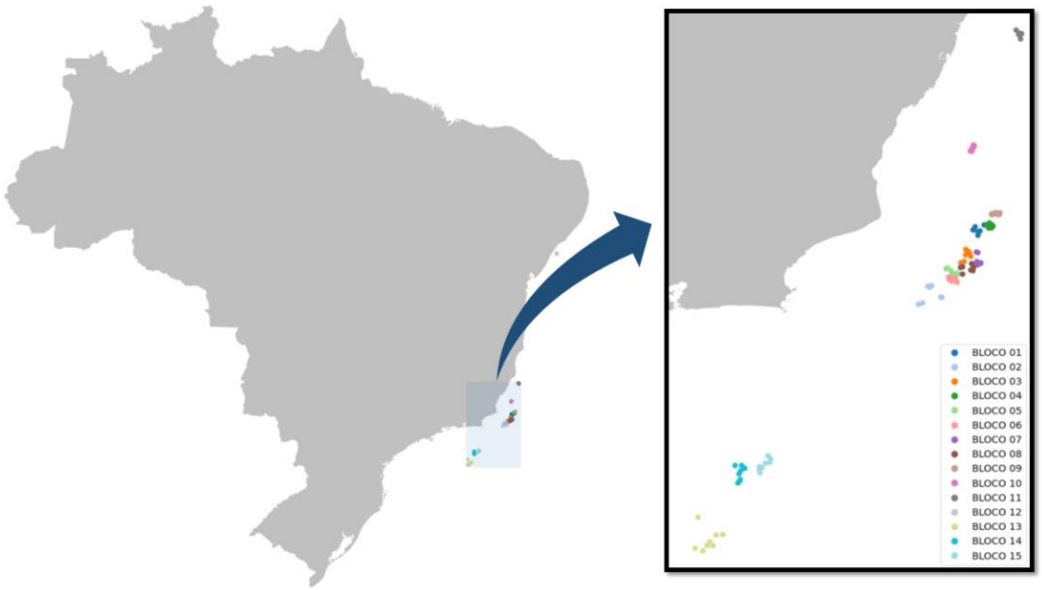


Figura 34 - Mapa com os poços utilizados para o experimento

4.2
Geração da matriz de compatibilidade entre rochas artificialmente

Em um ambiente produtivo, a matriz de compatibilidade entre os tipos de rocha deverá ser preenchida por geólogos experientes. Entretanto, no contexto de experimentação deste trabalho, foi criado um algoritmo simples para gerar artificialmente valores de compatibilidade entre as rochas, de modo a permitir algumas análises relacionadas ao cálculo do fator de compatibilidade entre poços.

Para duas rochas r_1 e r_2 quaisquer, o valor de compatibilidade artificial foi calculado como a razão entre a quantidade de poços em que ambas as rochas existem no perfil litológico e a quantidade de poços em que pelo menos uma das duas rochas existe no perfil litológico. A Figura 35 abaixo mostra o algoritmo criado para a geração desta matriz artificialmente e a Figura 36 em seguida mostra o mapa de calor (*heatmap*) da matriz gerada, contendo os valores resultantes.

Algoritmo 4: Geração artificial da matriz de compatibilidade entre rochas

Entrada:

D: dataset com os perfis litológicos de todos os poços

$rochas$: vetor contendo n tipos de rochas

Saída: Matriz de dimensões $n \times n$ contendo a compatibilidade para cada par de rochas

Método:

matriz = nova matriz de dimensões $n \times n$
Para i de 0 até $len(rochas) - 1$:
    $r_1 = rochas[i]$
    Para j de $i + 1$ até $len(rochas) - 1$:
        $r_2 = rochas[j]$
        $poços\_possuem\_r1\_ou\_r2 =$ quantidade de poços em D que possuam r_1 ou r_2
        $poços\_possuem\_r1\_e\_r2 =$ quantidade de poços em D que possuam r_1 e r_2
        $\text{compatibilidade} = \frac{poços\_possuem\_r1\_e\_r2}{poços\_possuem\_r1\_ou\_r2}$
        $matriz[i][j] = compatibilidade$
        $matriz[j][i] = compatibilidade$

Figura 35 - Algoritmo de geração artificial de compatibilidades entre rochas

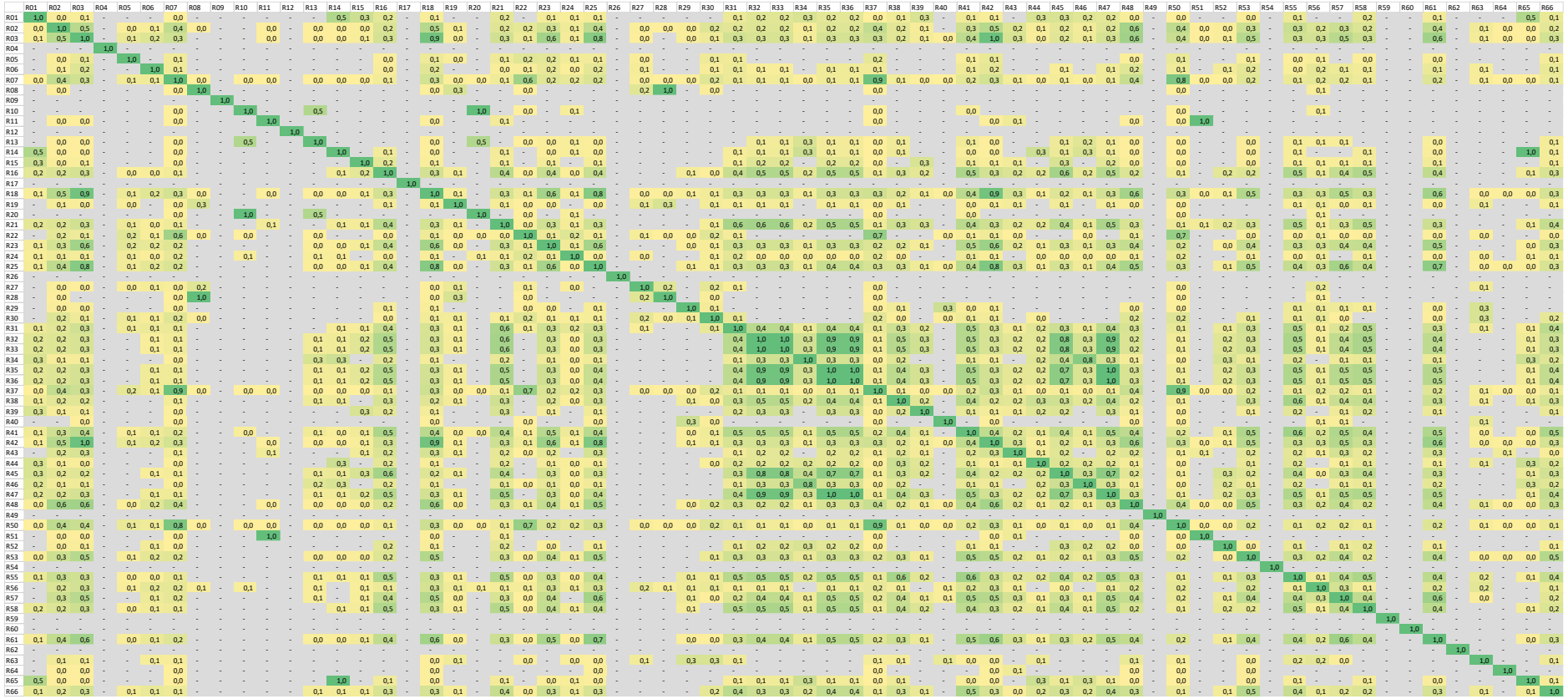


Figura 36 - Matriz de compatibilidades entre rochas gerada artificialmente para os experimentos do trabalho

4.3
Experimento

De acordo com o descrito nos três passos iniciais da metodologia, inicialmente foi gerado o **DNA Litológico** para cada um dos 120 poços coletados de acordo com a categorização de intervalos e identificação numérica das rochas. Na Figura 37 abaixo pode ser visualizada uma amostra dos dados gerados. Para a realização do quarto passo da metodologia foi implementado o algoritmo Needleman-Wunsch adaptado conforme proposto neste trabalho. Toda codificação necessária para a implementação da metodologia proposta neste trabalho foi realizada utilizando a versão 3 da linguagem Python (ROSSUM, 1995). Também foram usados os pacotes Pandas (MCKINNEY, 2010) e NumPy (HARRIS, MILLMAN, *et al.*, 2020) para tratamento de dados, o pacote PyClustering (NOVIKOV, 2019) para realização da clusterização através do K-Medoids, além do pacote Matplotlib (HUNTER, 2007) para geração de gráficos.

	ID_POCO	BLOCO	DNA
0	864396	BLOCO 01	0 1 1 1 2 1 3 1 4 49 .. 1032 26 1033 21 1034 21 1035 26 1036 21 Length: 1037, dtype: int64
1	839882	BLOCO 01	0 1 1 1 2 1 3 6 4 6 .. 779 26 780 21 781 26 782 21 783 21 Length: 784, dtype: int64
2	796789	BLOCO 03	0 1 1 1 2 1 3 1 4 36 .. 910 36 911 49 912 36 913 36 914 36 Length: 915, dtype: int64
3	94465	BLOCO 03	0 1 1 1 2 1 3 1 4 36 .. 792 36 793 36 794 49 795 36 796 36 Length: 797, dtype: int64
4	475698	BLOCO 03	0 1 1 1 2 1 3 1 4 6 .. 877 36 878 49 879 36 880 36 881 36 Length: 882, dtype: int64
5	202133	BLOCO 02	0 1 1 1 2 1 3 1 4 36 .. 830 6 831 49 832 49 833 49 834 49 Length: 835, dtype: int64
6	986572	BLOCO 06	0 1 1 1 2 1 3 1 4 1 .. 537 36 538 21 539 49 540 49 541 49 Length: 542, dtype: int64
7	283119	BLOCO 02	0 1 1 1 2 1 3 1 4 36 .. 1041 2 1042 2 1043 41 1044 41 1045 41 Length: 1046, dtype: int64
8	357002	BLOCO 02	0 1 1 1 2 1 3 1 4 1 .. 697 49 698 49 699 36 700 36 701 36 Length: 702, dtype: int64
9	953994	BLOCO 01	0 36 1 49 2 36 3 49 4 36 .. 727 6 728 6 729 21 730 21 731 21 Length: 732, dtype: int64
10	89047	BLOCO 05	0 6 1 36 2 6 3 6 4 6 .. 504 36 505 36 506 6 507 49 508 36 Length: 509, dtype: int64
11	842693	BLOCO 03	0 36 1 6 2 36 3 21 4 36 .. 610 36 611 49 612 49 613 36 614 49 Length: 615, dtype: int64

Figura 37 - Poços com DNAs Litológicos gerados

Em seguida o quarto passo da metodologia proposta foi realizado, através da reimplementação do algoritmo Needleman-Wunsch, incluindo os algoritmos para obtenção de similaridade e compatibilidades (por média ponderada, mínimo ou máximo) dos erros.

Para o passo seguinte da metodologia, o quinto, foram criadas 4 matrizes de dimensões 120×120, para armazenarem os valores de similaridade, compatibilidade por média ponderada, por mínimo e por máximo, para todos os

pares de poços. Essas matrizes foram armazenadas em arquivos CSV (*Comma Separated Values*) para que pudessem ser carregadas a qualquer momento. Este passo foi o mais custoso computacionalmente, tendo levado mais de 20 horas para a total execução em um computador científico básico, com processador Intel® Core™ i7 e 16GB de memória RAM.

Com todas as matrizes geradas, primeiramente foram gerados os mapas de calor ordenando os poços pelos blocos, permitindo, assim, uma visualização das possíveis correlações entre blocos. As imagens a seguir mostram esses mapas, com a coloração indo do amarelo (valores mais próximos de zero) ao verde (valores mais próximos de um).

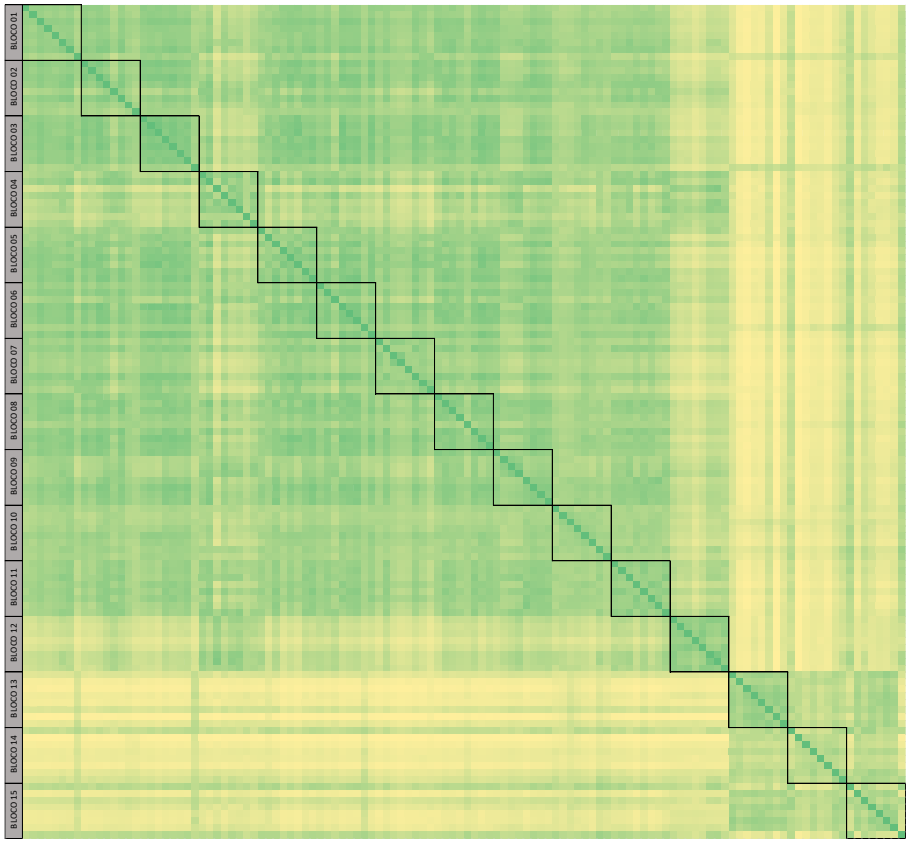


Figura 38 - Matriz de similaridades entre os poços

Pode ser verificado, no exemplo da Figura 38 acima, que os blocos 13, 14 e 15 são similares entre si, assim como os blocos 1 a 3 e 5 a 11 possuem alta similaridade entre si. Como as compatibilidades entre rochas, geradas artificialmente, levaram em consideração a frequência com que cada rocha está presente nos perfis litológicos dos poços, as matrizes de compatibilidade entre

poços acabaram por refletir uma equivalência à matriz de similaridades entre poços, tendo cada tipo de cálculo de compatibilidade (média ponderada, peso máximo e peso mínimo) se transformado em um grau de suavização da própria similaridade, como pode ser visto nas figuras 39, 40 e 41 abaixo.

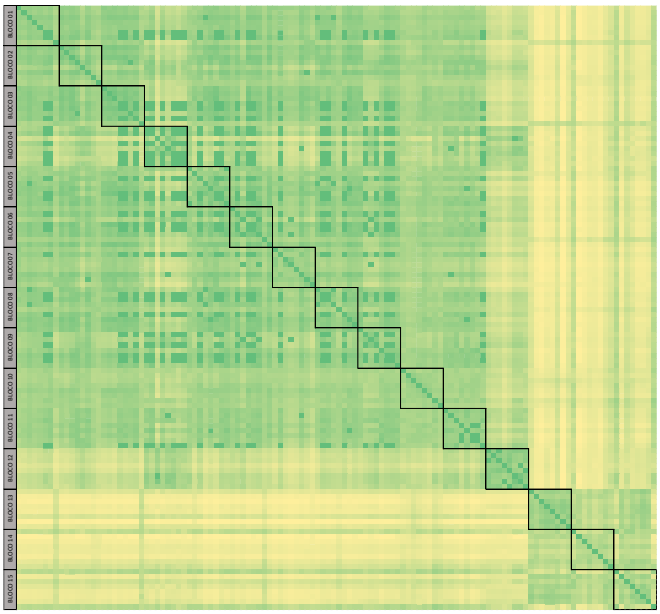


Figura 39 – Matriz de compatibilidades entre os poços pela média ponderada dos pesos

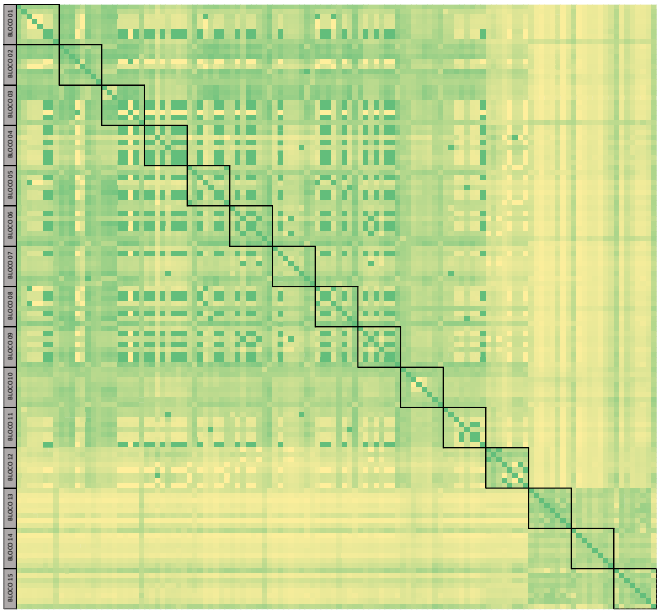


Figura 40 - Matriz de compatibilidades entre os poços pelo peso máximo

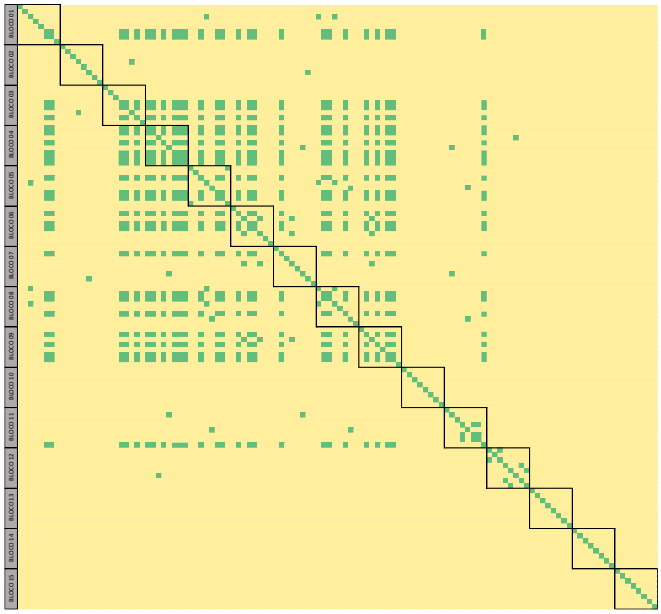


Figura 41 - Matriz de compatibilidades entre os poços pelo peso mínimo

O próximo passo do experimento foi calcular o número ideal de *clusters*, identificado pela variável k. Para isso, foi utilizado o método de cotovelo (*Elbow Method*), através das distâncias obtidas pela subtração da matriz de similaridades do valor 1 (um). O gráfico apresentado na Figura 42 abaixo mostra que $k = 10$ é o valor ideal para o *dataset* em questão.

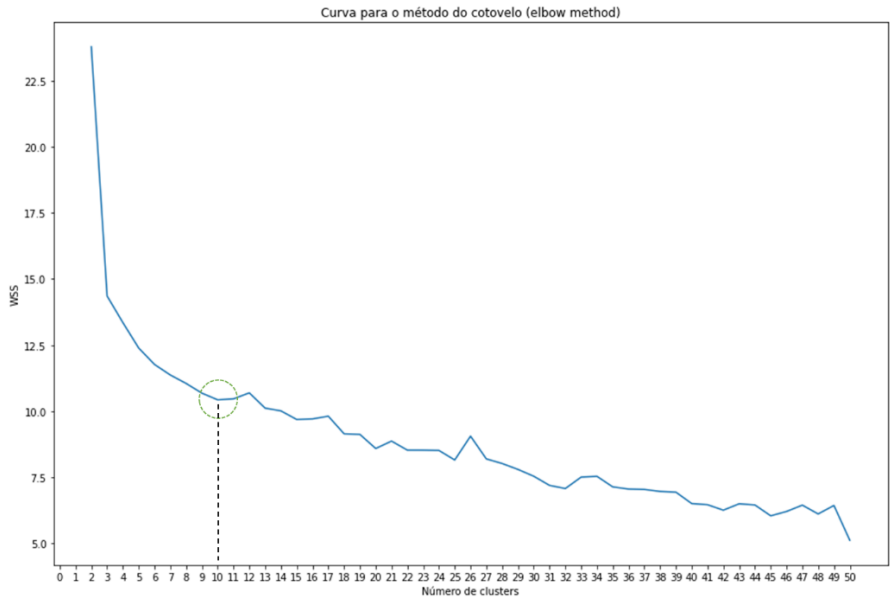


Figura 42 - Gráfico do método do cotovelo para obtenção do k ideal

Por fim foi executada a clusterização dos poços através do algoritmo K-medoids utilizando a matriz de distâncias (gerada a partir da matriz de similaridade) e $k = 10$ como parâmetros. Na Figura 43 abaixo pode ser observado o resultado da clusterização, mostrando que os *cluters* foram formados por poços independentemente do bloco de origem, já que foram agrupados pela similaridade de seus perfis litológicos e não por localização geográfica.

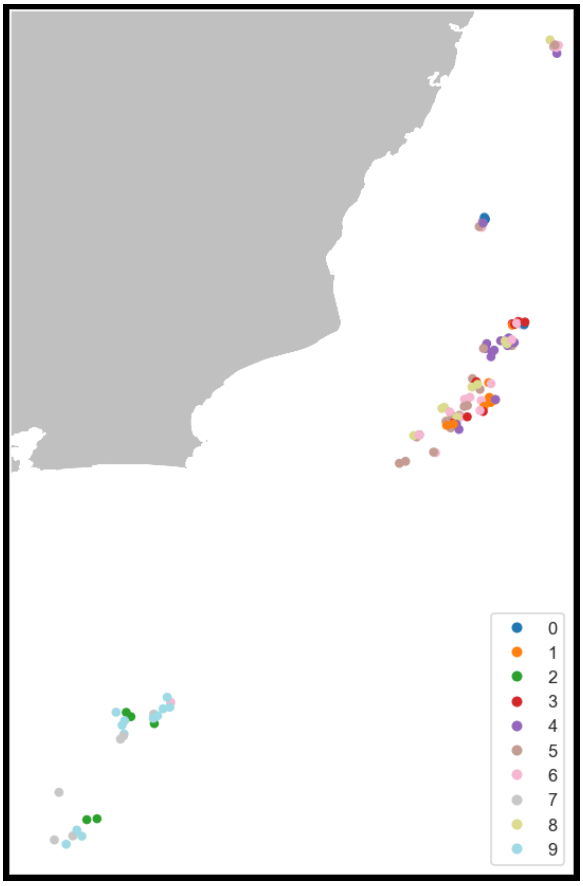


Figura 43 - Poços no mapa após clusterização

4.4 Resultados

A título de exemplificação, na Figura 44 abaixo estão dispostos perfis litológicos de 15 poços, um de cada um dos 15 blocos analisados, permitindo, assim, que se possa ter uma noção básica de como as regiões são diferentes entre si. Para isto foi também implementado um algoritmo para imprimir os perfis litológicos de poços, que foi utilizado para demonstrar os resultados desta pesquisa.

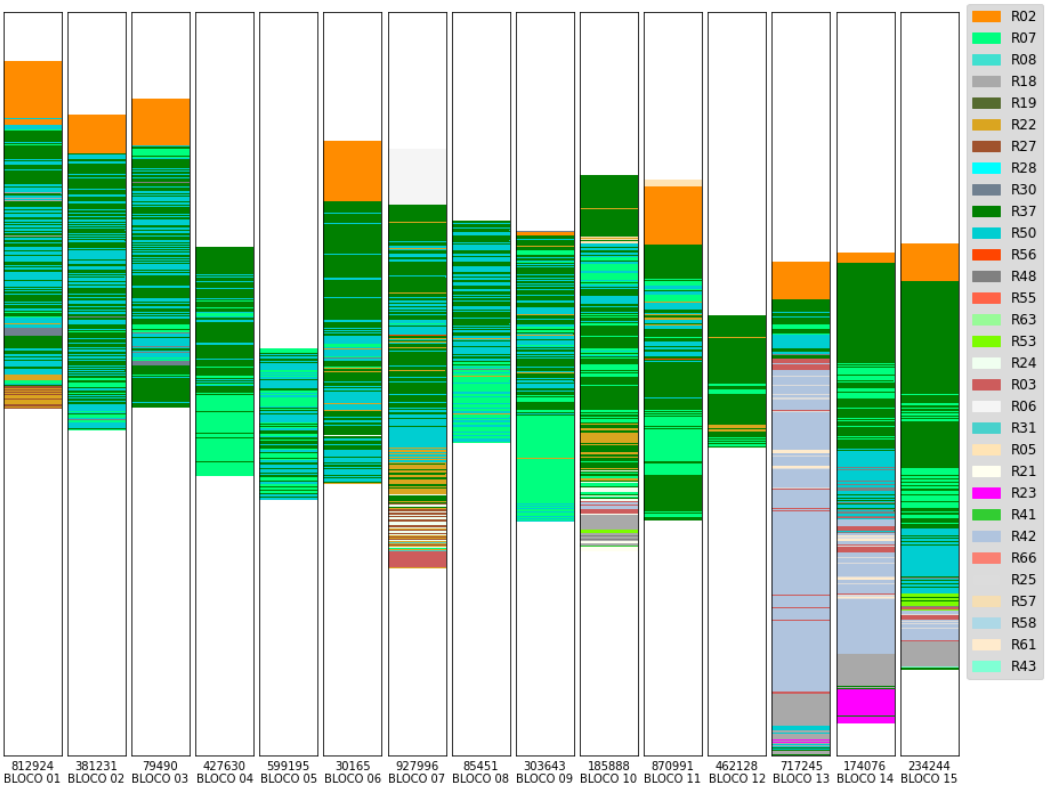


Figura 44 - Perfis litológicos de poços dos 15 blocos analisados

Após realizada a clusterização, o mapa de calor de similaridades foi gerado e é mostrado na Figura 45 a seguir. Nele pode ser observado que cada cluster gerado possui elementos similares entre si.

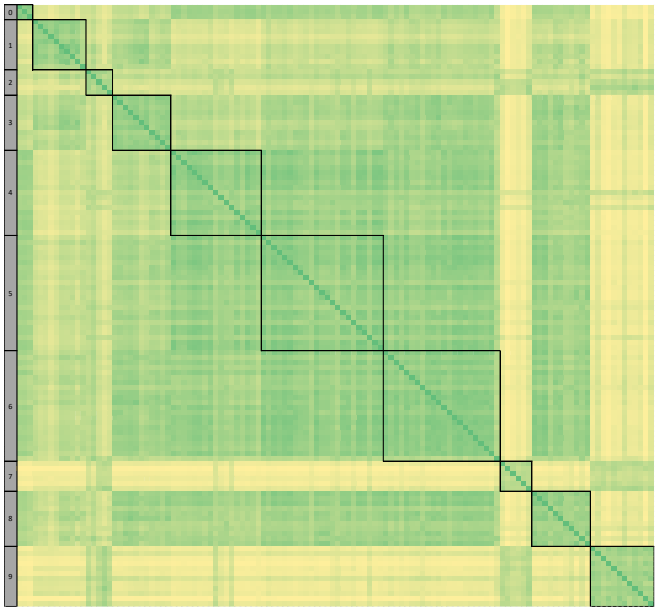
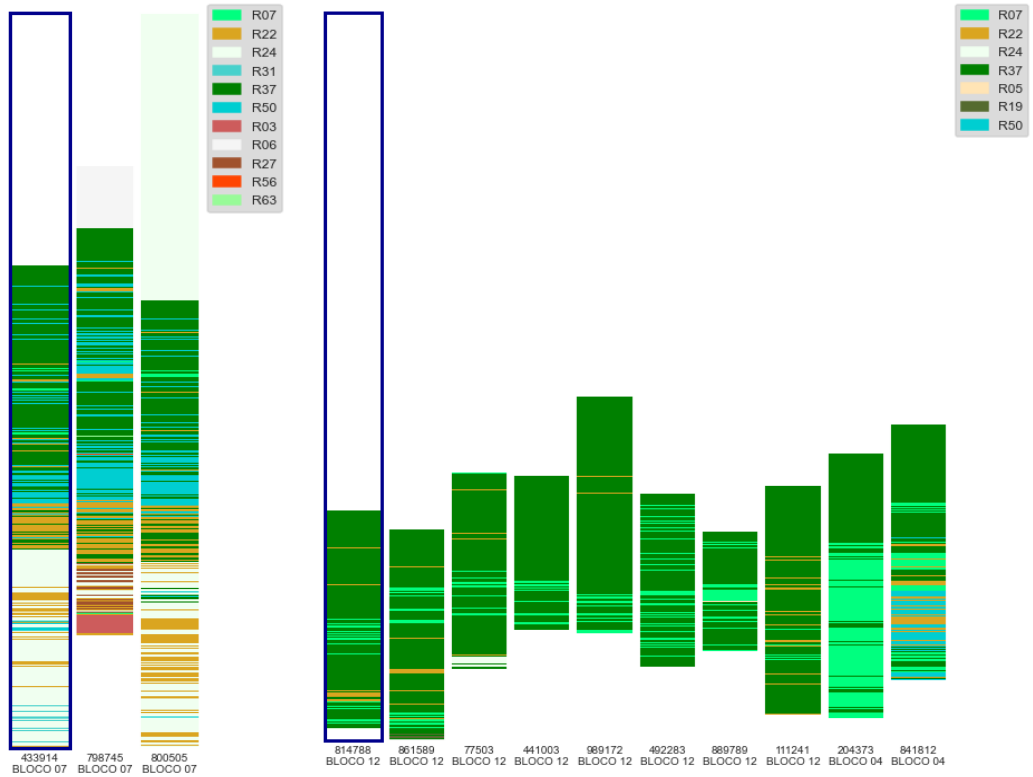


Figura 45 - Mapa de calor de similaridades entre poços agrupados por clusters

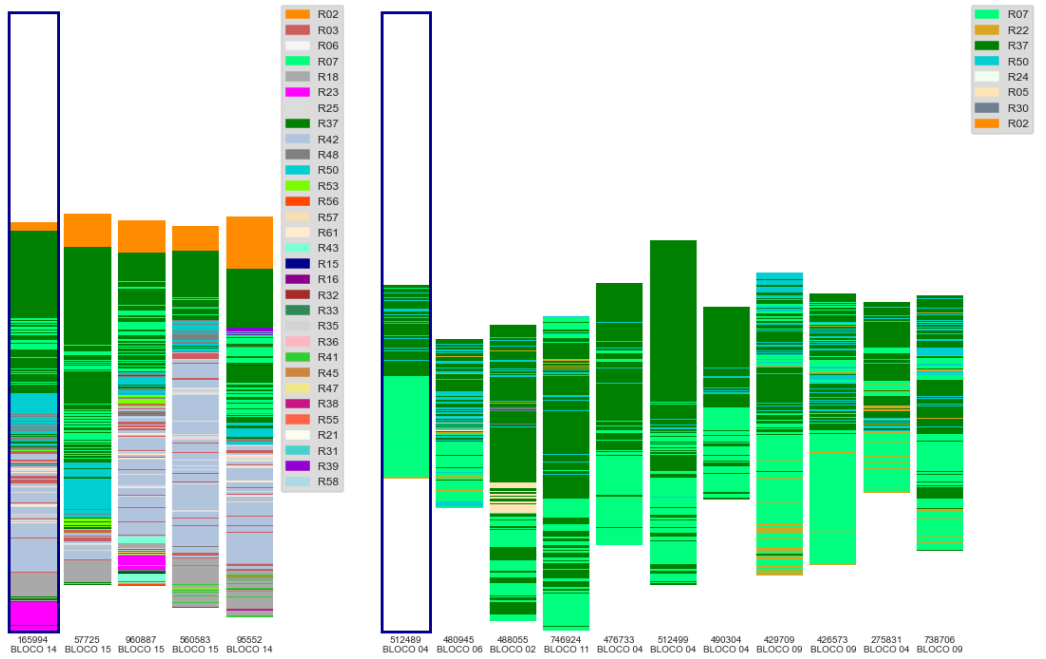
Em seguida, foram geradas impressões de comparações dos perfis litológicos dos poços para os 10 *clusters* gerados. Estas impressões serviram para uma primeira validação visual do resultado da clusterização utilizando a metodologia proposta. Os poços marcados em azul (que estão sempre no início de cada imagem) são os *medoids* de cada *cluster*.



(a) (b)

Figura 46 - Clusters 0 e 1 gerados

Na Figura 46 acima pode ser notado que (a) o *cluster* 0 possui apenas poços do bloco 7, entretanto, (b) o *cluster* 1 possui os 8 poços do bloco 12 e mais 2 poços do bloco 4. Pode ser notado, também, que os poços do bloco 7 classificados no cluster 0 possuem rochas muito características e similares. Já a Figura 47 abaixo mostra os *clusters* 2 e 3, tendo (a) o primeiro deles 5 poços dos bloco 14 e 15 e, (b) o segundo contendo poços de diversos blocos, mas com grande semelhança entre os perfis litológicos.



(a) (b)

Figura 47 - Clusters 2 e 3 gerados

A Figura 48 abaixo mostra a representação do cluster de número 4, contendo poços de diversos blocos, todos eles possuindo uma grande parcela de rochas semelhantes.

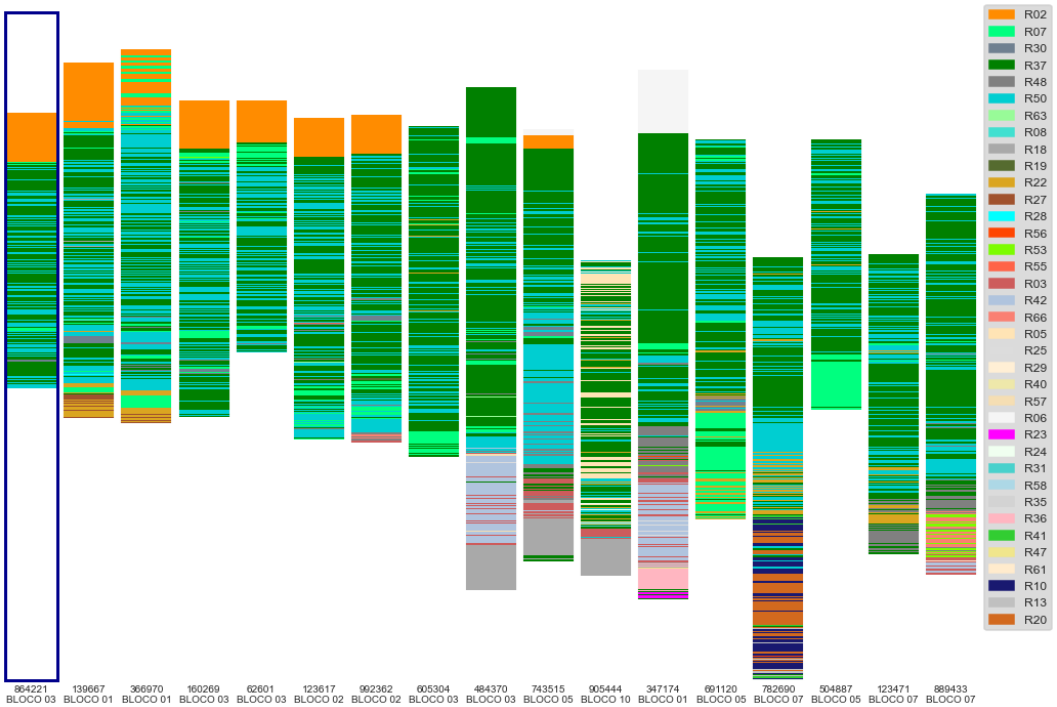


Figura 48 - Cluster 4 gerado

O mesmo pode ser observado nas figuras 49 e 50 abaixo, que mostram os clusters 5 e 6 gerados pela metodologia com base nos 120 poços disponibilizados no *dataset*.

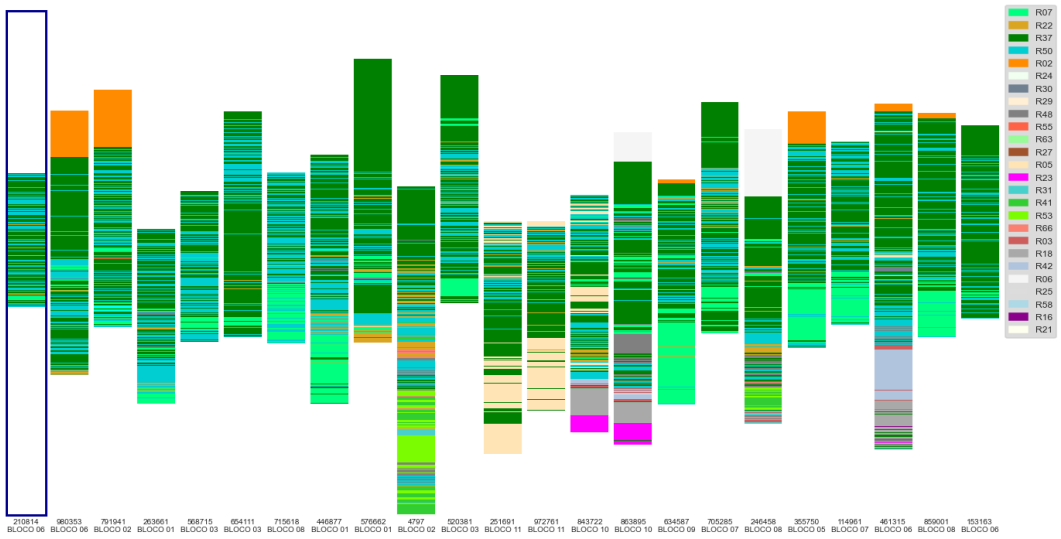


Figura 49 - Cluster 5 gerado

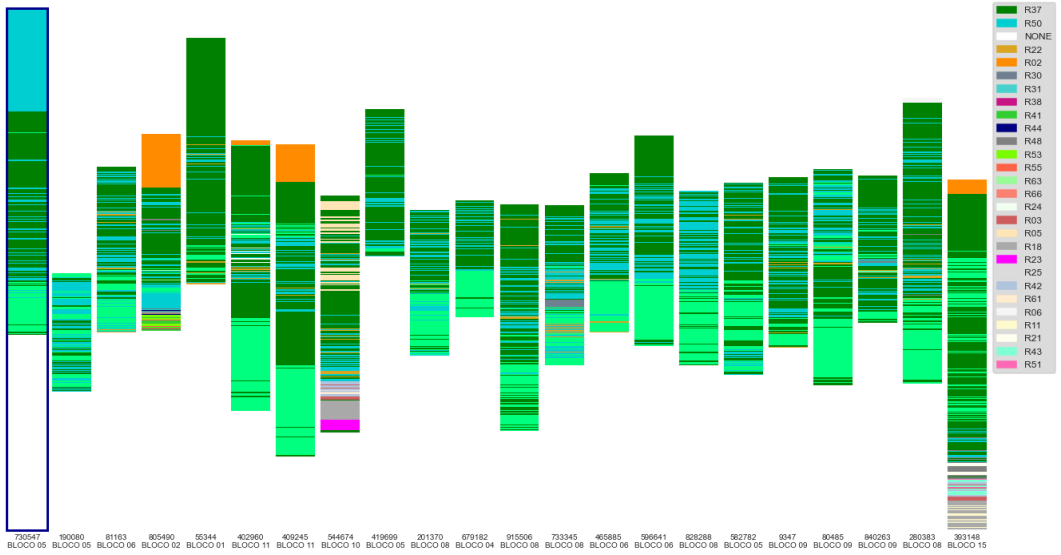
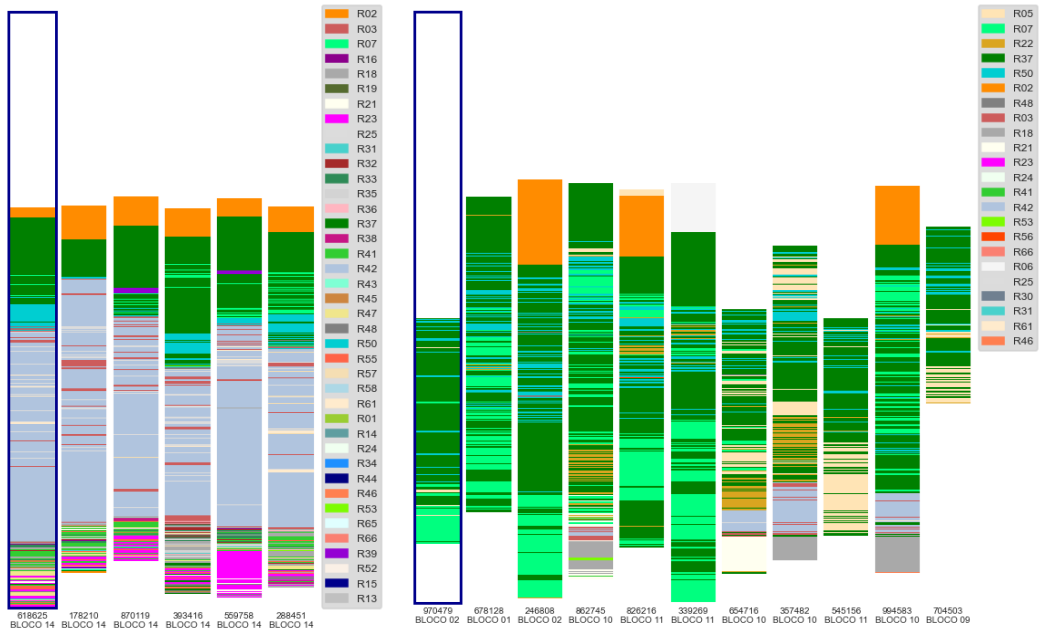


Figura 50 - Cluster 6 gerado

Na Figura 51 abaixo estão representados os clusters 7 e 8 gerados. Também pode ser verificada a semelhança entre os perfis litológicos dos poços que compõem o mesmo *cluster*.



(a) (b)

Figura 51 - Clusters 7 e 8 gerados

Finalmente o *cluster* 9 é apresentado na Figura 52 abaixo. Novamente é possível notar a semelhança entre os perfis litológicos dos poços que foram classificados neste *cluster*.

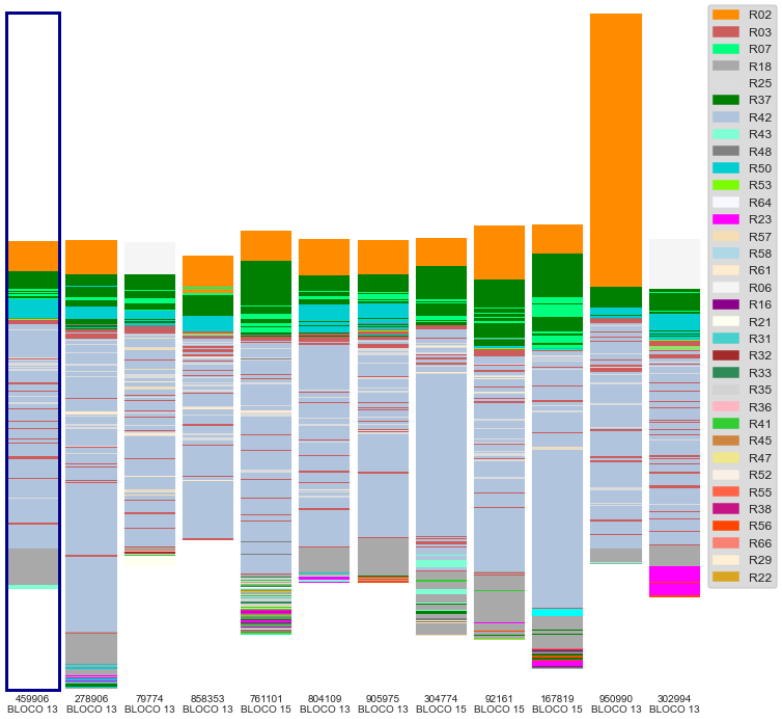
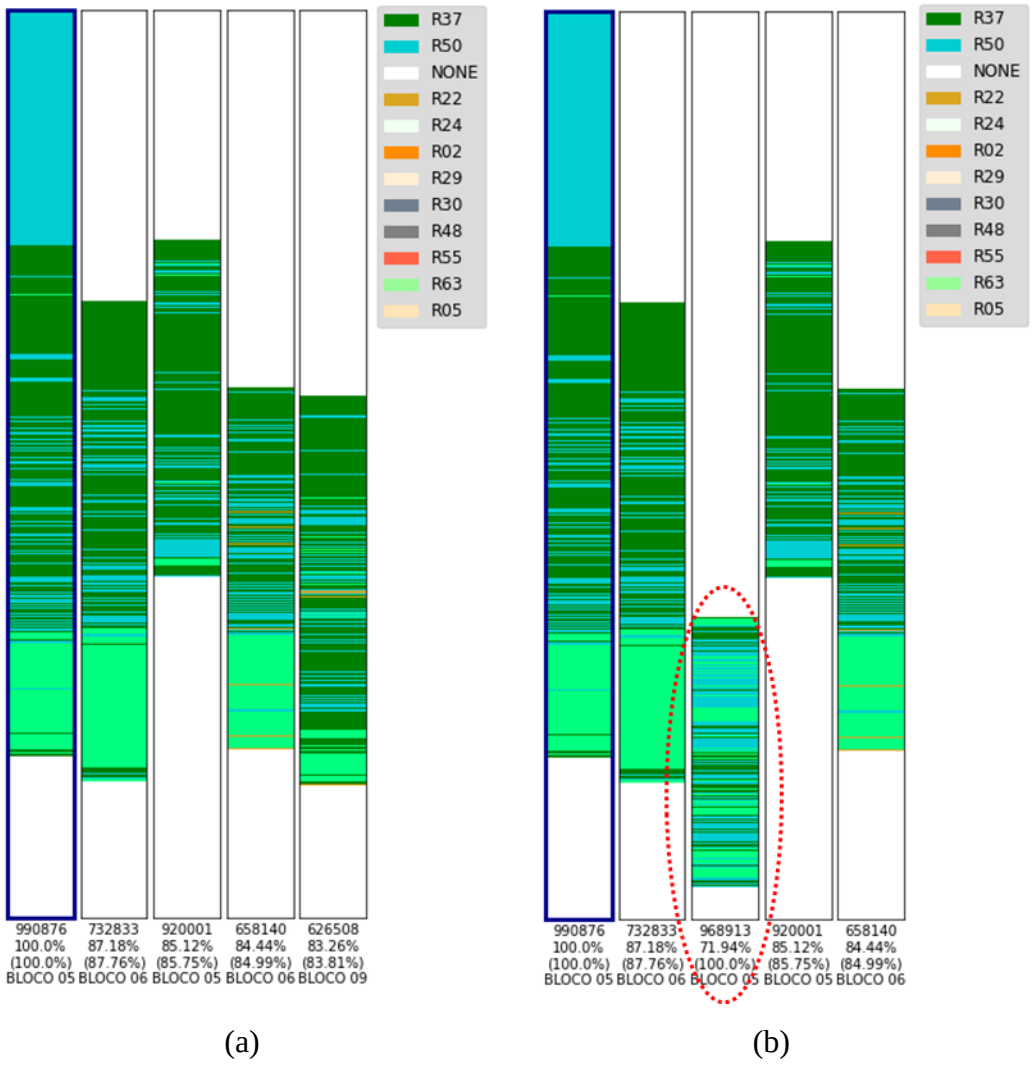


Figura 52 - Cluster 9 gerado

Além da análise por *clusters* gerados através da clusterização por similaridades, também foram geradas comparações entre poços utilizando as matrizes de compatibilidade entre poços. Apesar da matriz de compatibilidade entre rochas ter sido gerada de maneira artificial, foi possível avaliar a importância dos fatores de compatibilidade durante a comparação entre poços. Alguns exemplos são mostrados a seguir.



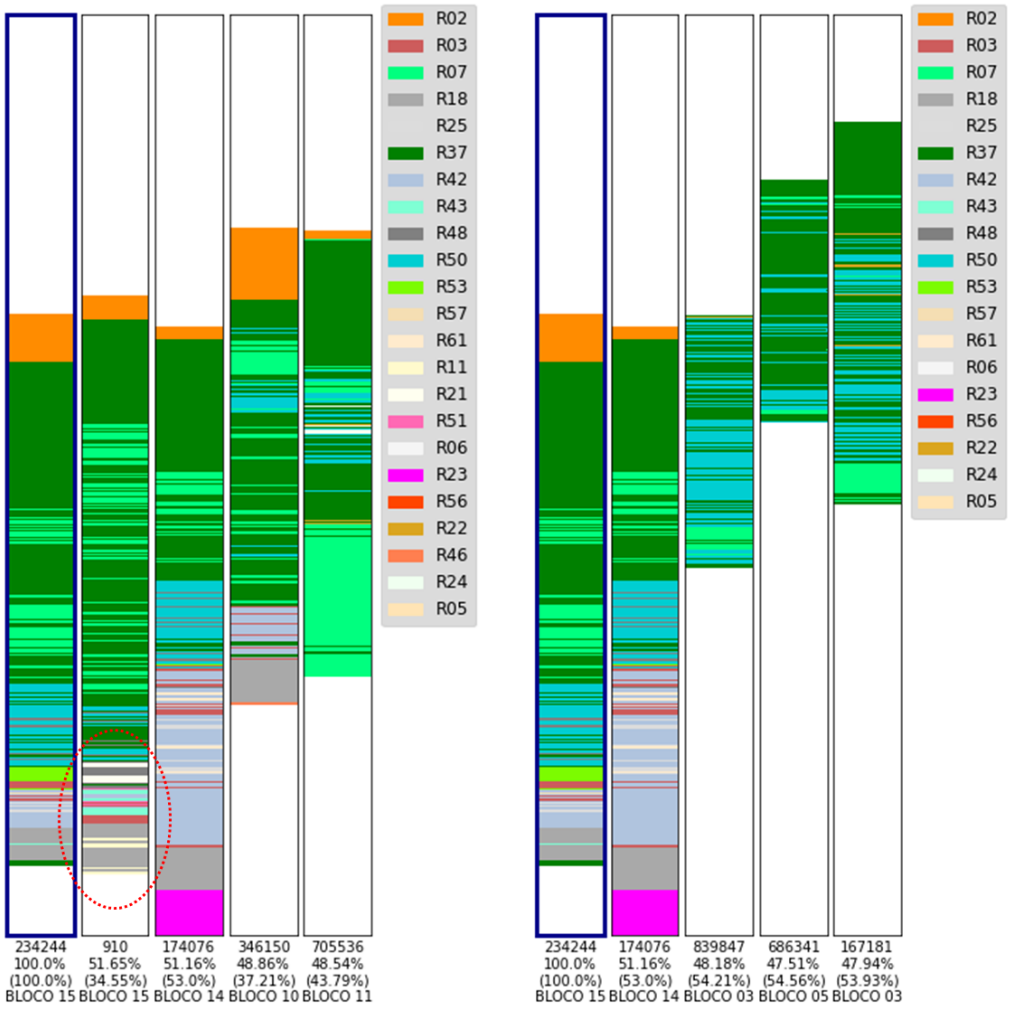
(a) (b)

Figura 53 - Influência da compatibilidade por média ponderada entre poços com rochas semelhantes

A Figura 53 acima mostra um determinado poço (de código 990876) e (a) os 4 poços mais similares a ele, ordenados apenas pela distância (grau de similaridade) entre eles. Já em (b) o critério de ordenação foi a média entre similaridade e compatibilidade para cada par de poços. Neste caso, nota-se uma

mudança na ordenação dos poços similares, ficando o poço de código 968913 (marcado em vermelho) com a segunda posição na sequência de poços similares, pois sua compatibilidade é de 100% com o poço estudado. Isto ocorreu pois não há nenhuma rocha exclusiva nestes dois poços.

Já a Figura 54 abaixo mostra um exemplo contrário, onde o poço base (marcado em azul) possui muita similaridade com o poço de código 910, como pode ser visto em (a), mas por este poço possuir rochas não existentes no poço base (marcadas em vermelho), sua compatibilidade por média ponderada se torna baixa (apenas 34,55%). Quando os poços são listados por ordem de produto entre similaridade e compatibilidade, este poço deixa de constar na lista de poços semelhantes, como pode ser visto em (b).



(a) (b)

Figura 54 - Influência da compatibilidade por média ponderada entre poços com rochas diferentes

Quando a compatibilidade por peso máximo é verificada, nota-se que algumas rochas, quando presentes em apenas um dos poços do par analisado, pode impactar substancialmente no valor final de compatibilidade entre dois poços. Este comportamento é esperado de acordo com o que foi proposto na metodologia. Um exemplo pode ser visto nas figuras 55 e 56 abaixo. Nelas pode ser analisada a sequência de poços similares ao poço 774234.

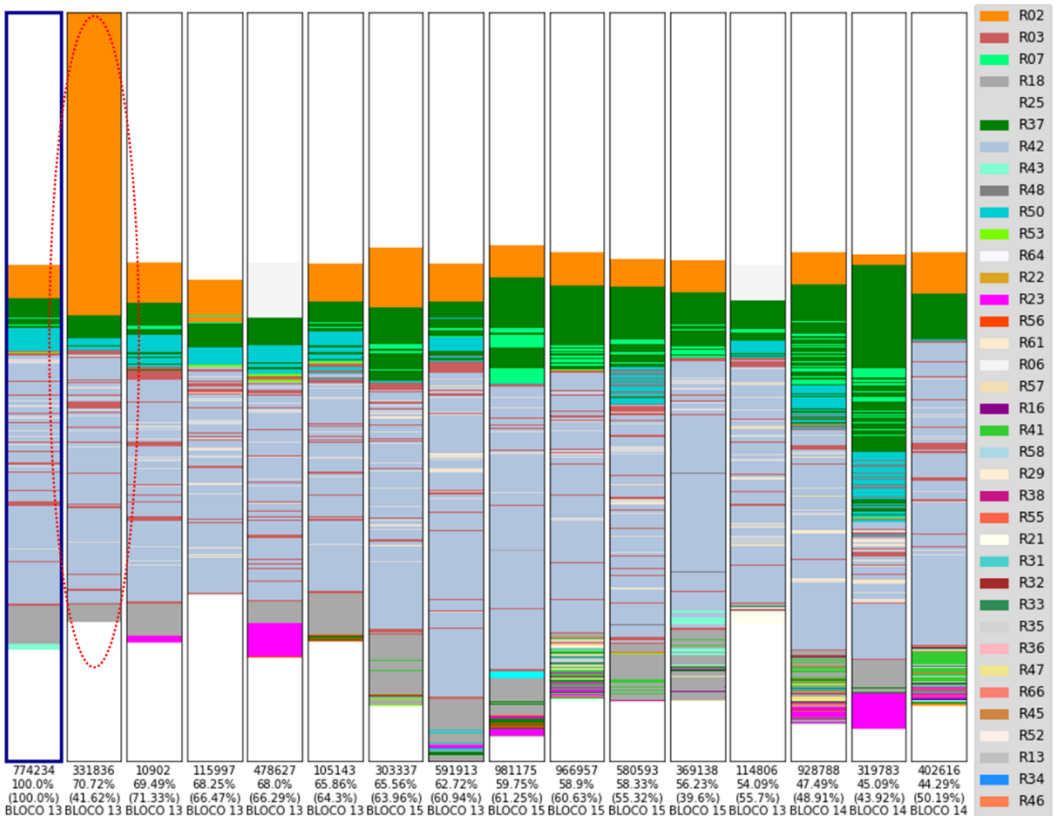


Figura 55 - Poços similares ao poço 774234 ordenados por similaridade

Na Figura 55 acima o poço 331836 (marcado em vermelho), chamado de "Poço B", pode ser visto como o mais similar ao poço estudado (de código 774234), chamado de "Poço A". Porém, como existem rochas presentes no "Poço A" não estão presentes no "Poço B" e, adicionado a isso, a compatibilidade atribuída a estas rochas em relação às rochas existentes no "Poço B" é baixa, temos que a compatibilidade máxima entre o "Poço A" e o "Poço B" tenha sido calculada como baixa (apenas 41,62%). Isto faz com que, ao serem listados os poços semelhantes ao "Poço A" ordenados pelo produto entre a similaridade e a compatibilidade por peso máximo, o "Poço B" sai da primeira posição e só aparece na 11ª posição, como

pode ser visto na Figura 56 abaixo. Nela o “Poço B” também está marcado com uma linha tracejada de cor vermelha.

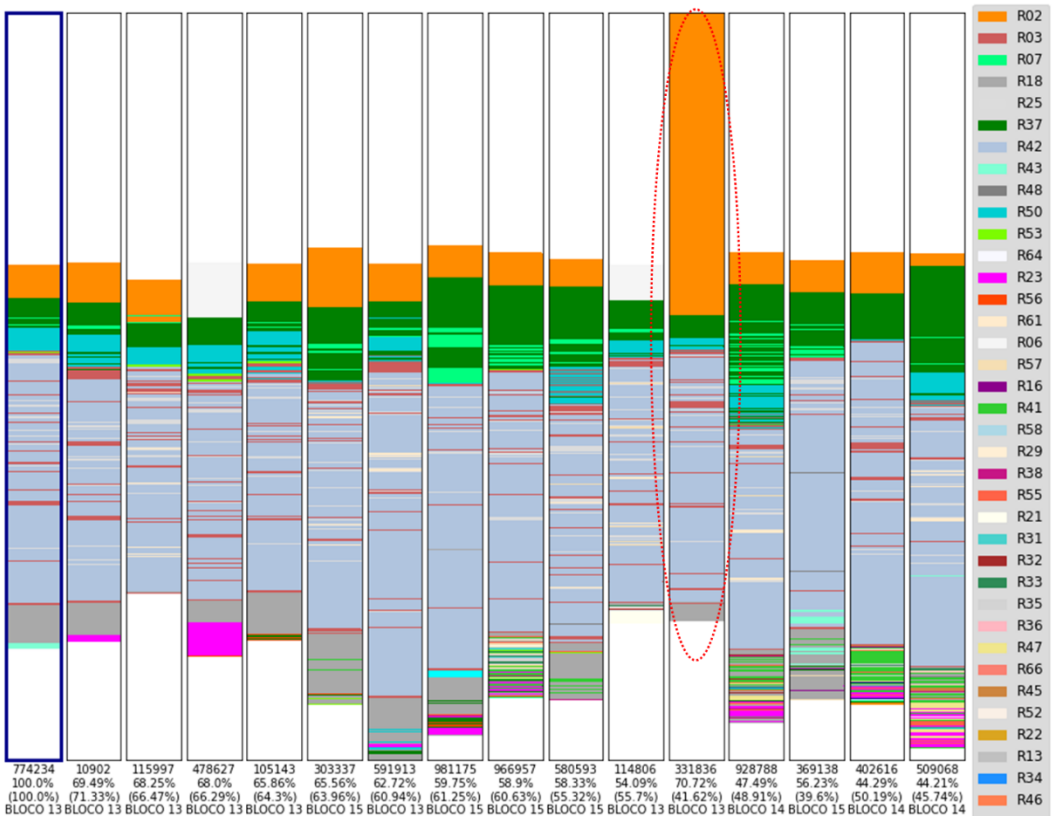


Figura 56 - Poços similares ao poço 774234 ordenados pelo produto entre similaridade e compatibilidade por peso máximo

Como a matriz de compatibilidades entre rochas foi gerada artificialmente, o método utilizado para gerar essas compatibilidades fez com que os valores encontrados como resultado do cálculo de compatibilidade por peso mínimo entre os poços refletissem apenas uma boa opção para poços com poucas rochas em seu perfil litológico, de modo que fossem removidos dos poços similares aqueles que não possuem as mesmas rochas que o poço estudado. Um exemplo pode ser visto através das figuras 57 e 58 a seguir. Na Figura 57 abaixo é mostrada a sequência de poços similares ao poço 920001. Os poços apontados com setas vermelhas possuem compatibilidade mínima igual a zero em relação ao poço objeto de estudo (de código 920001), pois possuem rochas não existentes no poço base de estudo. Ao listar os poços similares a este mesmo poço, mas ordenando pelo produto entre o grau de similaridade e a compatibilidade pelo peso mínimo, apenas poços que

possuem as mesmas rochas que o poço 92001 são listados (considerando os 15 mais similares da imagem), como pode ser visto na Figura 58 mais abaixo.

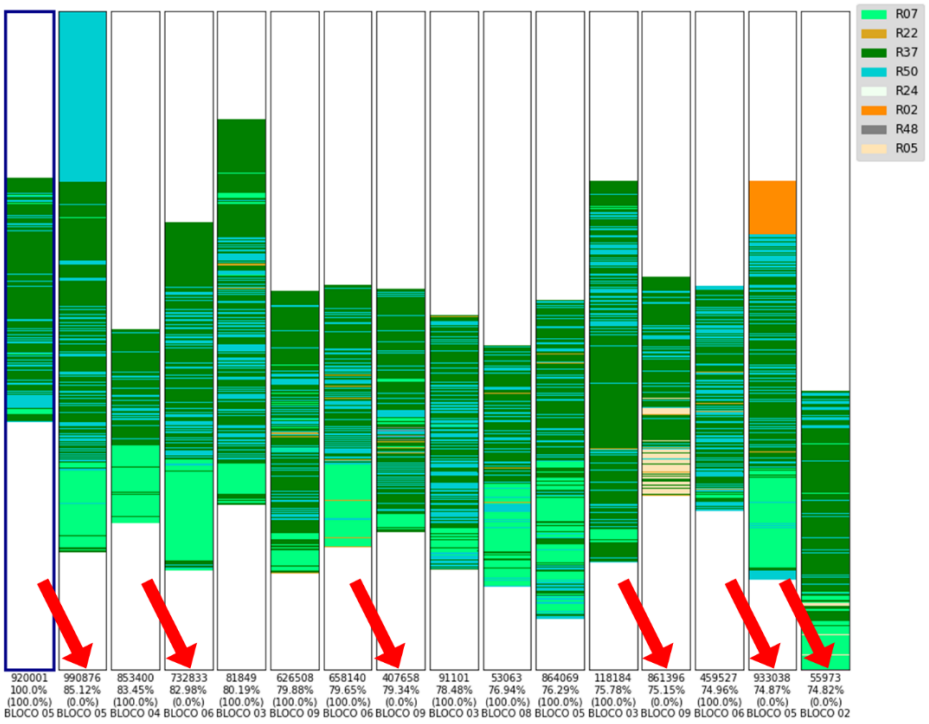


Figura 57 - Poços similares ao poço 920001 ordenados por similaridade

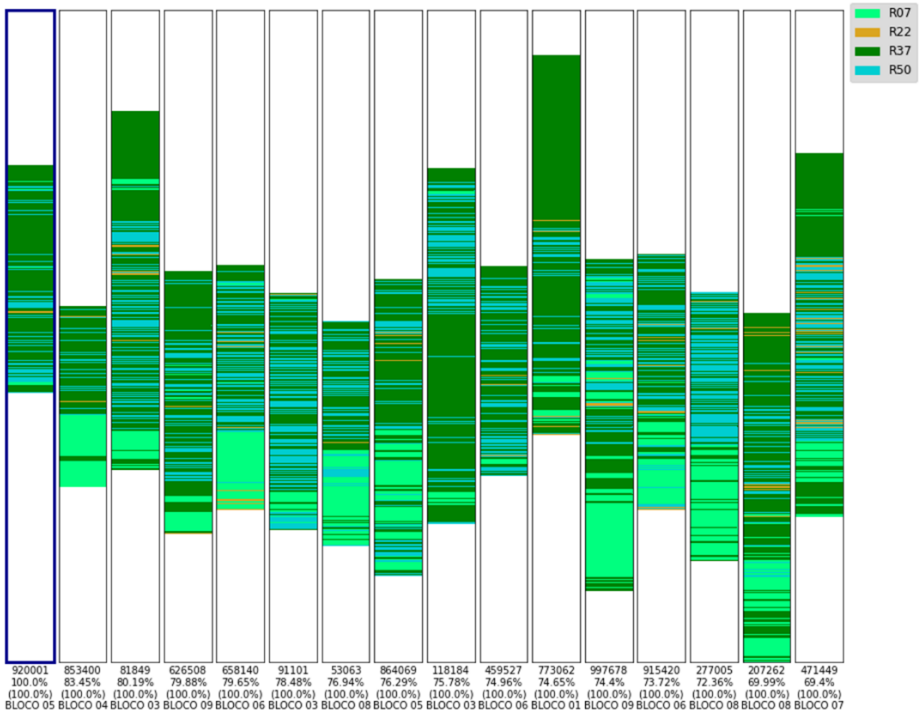


Figura 58 - Poços similares ao poço 920001 ordenados pelo produto entre similaridade e compatibilidade por peso mínimo

5
Considerações Finais

Esta dissertação propôs uma metodologia para clusterização de poços de petróleo através do método de particionamento baseado em medoids, através do algoritmo K-medoids. Como entrada para o algoritmo de clusterização, foi proposta a utilização de uma matriz de distâncias gerada por um algoritmo de correlação de poços de petróleo. O algoritmo de correlação primeiramente transforma o perfil litológico de um poço de petróleo em uma sequência denominada DNA Litológico do poço. A partir dos DNAs Litológicos dos poços, o algoritmo de alinhamento global de sequências Needleman-Wunsch foi adaptado para ser executado para cada par de poços e, então, retornar as medidas de similaridade e compatibilidade para cada par de poços. Finalmente, essa medida de similaridade foi então utilizada para a geração da matriz de distâncias e, assim, possibilitar a realização da clusterização.

Durante a revisão bibliográfica foi possível rever conceitos básicos sobre construção de poços de petróleo, litologia, alinhamento de sequências de DNA e clusterização de dados. Todos esses conceitos foram fundamentais para a construção da metodologia proposta. Até o momento da finalização deste trabalho não foi encontrado, na base SCOPUS, outro trabalho que abordasse o assunto de correlação de poços de petróleo através de algoritmos de alinhamento de sequências baseadas exclusivamente na litologia dos poços. Alguns trabalhos abordam esses assuntos, mas de maneiras distintas. Estes trabalhos também foram discutidos nesta dissertação, assim como as lacunas neles existentes e quais seriam as soluções propostas. Esses aspectos foram considerados para a definição da metodologia proposta.

Os experimentos realizados, através dos dados de perfis litológicos de 120 poços de petróleo da costa sudeste do Brasil, mostraram que a metodologia proposta atingiu um resultado muito satisfatório com relação à clusterização dos poços com base em seus perfis litológicos, ou seja, a medida de distância calculada entre poços pôde ser observada como adequada. No entanto, como a matriz de compatibilidade entre rochas não foi gerada por especialistas, mas sim através de uma construção

artificial baseada na amostragem de cada rocha no *dataset* de testes, a medida de compatibilidade gerada pela metodologia não teve muita relevância prática para os experimentos em si, mas pôde ser notado que, se corretamente parametrizada, poderá ser uma medida de suma importância no processo de correlação entre poços, principalmente pelo fato de poder ser preenchida com pesos (compatibilidades) diferentes de acordo com o objetivo de cada análise.

A conclusão deste trabalho é que a correlação de poços utilizando os dados de perfis litológicos, através de algoritmos de alinhamento global de sequências, é uma forma viável e, principalmente, independente de análise de especialistas, podendo ser utilizada como uma primeira triagem para uma série de validações automáticas, como checagens de possíveis inconsistências nas definições de materiais e equipamentos alocados para construção de novos poços de petróleo.

5.1
Sugestões para trabalhos futuros

Como os identificadores dos poços foram descaracterizados, não foi possível gerar métricas que permitissem uma melhor validação da metodologia proposta. A análise foi realizada de maneira visual através de diversos gráficos gerados durante os experimentos. Um novo tipo de análise que pode ser realizado é a consulta a dados de outras bases de dados, a fim de verificar se poços pertencentes ao mesmo *cluster* possuem características semelhantes, como, por exemplo, materiais ou equipamentos utilizados para construção de poços. Com essas informações, será possível calibrar as categorias para cada intervalo de rochas mais adequadamente.

Outro aspecto que poderá ser abordado em trabalhos futuros é a utilização de dados de mais poços, incluindo poços terrestres e também de outras regiões, assim como a criação de um novo parâmetro para definição de grupos de rochas semelhantes, de modo que rochas que estejam no mesmo grupo passem a ser consideradas como iguais durante a execução do algoritmo de alinhamento de sequências, gerando *matches* em vez de *mismatches*.

Além disso, trabalhos futuros também poderão abordar aspectos relacionados ao desempenho dos algoritmos propostos, através de otimizações e modificações para serem paralelizados, permitindo a obtenção dos resultados mais rapidamente, principalmente quando a metodologia for empregada em um *dataset* maior.

Bibliografia

ARNOLD, L. Unsupervised Example: Clustering and K-means. **From Optimization to Machine Learning**, 13 jul. 2021. Disponivel em: <https://ludovicarnold.com/teaching/optimization-machine-learning/unsupervised-example-clustering-k-means/>. Acesso em: 12 jul. 2021.

ASSMANN, B. W. **Estudo de estratégias de otimização para poços de petroleo com elevação por bombeio de cavidades progressivas**. Natal, RN: UFRN, 2008.

BALDI, P.; BRUNAK, S. **Bioinformatics:** The Machine Learning Approach. [S.l.]: MIT Press, 1998.

BERKHIN, P. Survey of Clustering Data Mining Techniques. In: J., K.; C., N.; M., T. **Grouping Multidimensional Data**. Berlin: Springer, 2006. p. 25-71.

CORMEN, T. H.; LEISERSON, C. E.; RIVEST, R. L. **Introduction to Algorithms**. [S.l.]: MIT Press, 1989.

CRISTINO, A. D. S. Principais algoritmos de alinhamento de sequências genéticas. **CoteiaWiki**, 2012. Disponivel em: <http://wiki.icmc.usp.br/images/4/41/Scc0172_10.1_alinhamento_genetico.pdf>. Acesso em: 11 jul. 2021.

DARNELL, J. E.; LODISH, H. F.; BALTIMORE, D. **Molecular Cell Biology**. [S.l.]: Scientific Amer Inc., 1990.

DE SOUZA, R. G. **Alinhamento Múltiplo de Sequências Utilizando Otimização Dialética**. Recife, PE: UFPE, 2014.

FREIMANN, B. C.; ALVES, J. G. D. V.; SILVA, M. W. C. Estudo Hidrogeológico Através de Perfis Geofísicos de Poços – Salinópolis-PA. **Águas Subterrâneas**, 2014.

GAN, G.; MA, C.; WU, J. **Data clustering:** theory, algorithms and applications. Philadelphia, Pa: SIAM, Society for Industrial and Applied Mathematics, 2007.

GARCIA, L. F.; CARBONERA, J. L.; ABEL, M. **Ontologies applied to lithologic correlation problem within the petroleum geology domain**. 6th Seminar on Ontology Research in Brazil, ONTOBRAS 2013. [S.l.]: [s.n.]. 2013.

GIBBS, A. J.; MCINTYRE, G. A. The Diagram, a Method for Comparing Sequences, 1970.

GREEN, E. D. DNA Sequencing. **National Human Genome Research Institute**, 8 jul. 2021. Disponivel em: <https://www.genome.gov/genetics-glossary/DNA-Sequencing>.

HAIR JR., J. F. et al. **MULTIVARIATE DATA ANALYSIS - 7th Edition**. [S.l.]: Pearson, 2015.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining:** Concepts and Techniques. [S.l.]: Elsevier, 2012.

HARRIS, C. R. et al. Array programming with NumPy. **Nature**, p. 357–362, 2020.

HRUSCHKA, E. R.; EBECKEN, N. F. F. A genetic algorithm for cluster analysis. **Intelligent Data Analysis 7**, p. 15-25, 2003.

HUNTER, J. D. Matplotlib: A 2D Graphics Environment. **Computing in Science & Engineering**, p. 90-95, 2007.

IOSTE, A. R. **Sequências de DNA:** uma nova abordagem para o alinhamento ótimo. São Paulo, SP: PUC-SP, 2016.

ISSARANE, H. Le Clustering: Définition et Top 5 Algorithmes - Analytics & Insights. **Analytics & Insights**, 12 jul. 2021. Disponivel em: <https://analyticsinsights.io/le-clustering-definition-et-implementations/>. Acesso em: 15 jul. 2021.

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, 9 set. 2009.

JARVIS, K.; SAUSSUS, D. Extracting detailed lithology from seismic data. **ASEG Extended Abstracts**, 2009.

KAUFMAN, L.; ROUSSEEUW, P. J. Clustering by means of medoids, 1987.

KUMAR, V. **DATA CLUSTERING:** Algorithms and Applications. [S.l.]: CRC Press, 2014.

LEITE, V. R. C.; GATTASS, M. **Uma analise da classificação de litologias utilizando SVM, MLP e métodos Ensemble**. Rio de Janeiro, RJ: PUC-Rio, 2012.

MACQUEEN, J. **Some methods for classification and analysis of multivariate observations**. 5th Berkeley Symposium on Mathematical Statistics and Probability. [S.l.]: [s.n.]. 1967. p. 281-297.

MAHAJAN, M.; NIMBHORKAR, P.; VARADARAJAN, K. The planar k-means problem is NP-hard. **Theoretical Computer Science**, p. 13-21, 2012.

MARTINS, M. M.; HAMACHER, S.; ACCIOLY, R. D. M. E. S. **Comparação de estratégias de construção de poços marítimos incorporando incertezas**. [S.l.]: PUC-Rio, 2014.

MASCULO, M. S. Projeto de Construção de Poço. **SPETRO**, p. 3, 2012.

MATTA, M. A. D. S. et al. Geometria dos Sistemas Aqüíferos da Bacia Hidrográfica do Paracuri - Belém/PA, Como Base Para Uma Proposta de Abastecimento de Água Subterrânea. **XIII Congresso Brasileiro de Águas Subterrâneas**, 2004.

MATTOS, P. A. D. C. **Vibrações em Colunas de Perfuração em Operações na Bacia do Solimões**. Rio de Janeiro, RJ: PUC-Rio, 2015.

MATTT. DBSCAN. **NSHipster**, 13 fev. 2020. Disponivel em: <https://github.com/NSHipster/DBSCAN>.

MCKINNEY, W. Data Structures for Statistical Computing in Python. **Proceedings of the 9th Python in Science Conference**, p. 51-56, 2010.

NATIONAL HUMAN GENOME RESEARCH INSTITUTE. Human Genome Project FAQ. **National Human Genome Research Institute**, 20 jul. 2021. Disponivel em: <https://www.genome.gov/human-genome-project/Completion-FAQ>. Acesso em: 15 jul. 2021.

NAVER. k-medoids technique: representative object-based technique. **Naver Blog**, 18 mar. 2009. Disponivel em: <https://blog.naver.com/asus1984/120065317344>. Acesso em: 18 jul. 2021.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Jounal of Molecular Biology**, 28 mar. 1970.

NEVES, M. L. et al. **Alocação de Equipamentos Críticos em Projetos de Produção de Petróleo Offshore**. Rio de Janeiro, RJ: SBPO - Simpósio Brasileiro de Pesquisa Operacional, 2012.

NOVIKOV, A. PyClustering: Data Mining Library. **Journal of Open Source Software**, p. 1230, 2019.

OLIVEIRA, R. W. D.; GUIMARÃES, L. J. N.; MANZOLI, O. L. **Estudo da Integridade da Cimentação em Poços Submetidos à Injeção de Vapor via Técnica de Fragmentação de Malha de Elementos Finitos**. Recife: UFPE, 2016.

PELIANO, S. V.; NEGRÃO, C. O. R. **Análise experimental do arrasto em colunas de perfuração parcialmente imersas em leito de cascalhos**. [S.l.]: UTFPR, 2018.

PERETA, M. G. D. S. **Um Approach Evolucionário da Trajetória Tecnológica do Segmento Offshore na Indústria de Petróleo Brasileiro:** Os Desafios Tecnológicos do Pré-Sal. [S.l.]: Unicamp, 2015.

PETROBRAS. **Introdução à indústria do Petóleo**. [S.l.]: Petrobras, 2006.

PETROBRAS. Tecnologias pioneiras do PRÉ-SAL, 6 jul. 2021. Disponivel em: <https://presal.petrobras.com.br/tecnologias-pioneiras>.

PINHEIRO, L. N. P. **Caracterização do reservatório Carapebus do campo de Peregrino, bacia de Campos, através da análise de perfis geofísicos de poços, integrada à interpretação sísmica**. Niterói, RJ: UFF, 2014.

ROCHA, L. A. S.; DE AZEVEDO, C. T. **Projetos de Poços de Petróleo - Geopressões e Assentamento de Colunas de Revestimentos**. [S.l.]: Editora Interciência, 2019.

ROSA, A. J.; CARVALHO, R. D. S.; XAVIER, J. A. D. **Engenharia de Reservatórios de Petróleo**. [S.l.]: Editora Interciência, 2006.

ROSSUM, G. V. **Python tutorial, Technical Report**. Amsterdam. 1995.

SETUBAL, J.; MEIDANIS, J. **Introduction to computational molecular biology**. [S.l.]: PWS Publishing, 1997.

SMITH, T. F.; WATERMAN, M. S. Identification of Common Molecular Subsequences. **Journal of Molecular Biology**, 1981.

SOUZA, C. O. D. **Análise de Correlação Litológica a Partir de Dados de Perfis de Poços Convencionais do Campo de Namorado Usando Software Comercial**. Niterói, RJ: UFF, 2014.

THOMAS, J. E. **Fundamentos de Engenharia de Petróleo**. [S.l.]: Editora Interciência, 2004.

VILAS BOAS, D. B.; SOUZA, P. V.; HOLZ, M. **Correlação Sismoestratigráfica entre as bacias do Recôncavo e de Camamu**. São Paulo, SP: UNESP, 2018.