UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

LUAN FONSECA GARCIA

Uma Abordagem Automática para Extrair Correlações Geológicas a Partir de Descrições de Poços Baseadas em Ontologias.

Trabalho de Conclusão apresentado como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação

Profa. Dra. Mara Abel
Orientador

Msc. Joel Luis Carbonera
Co-orientador

Porto Alegre, dezembro de 2013

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Garcia, Luan Fonseca

Uma Abordagem Automática para Extrair Correlações Geológicas a Partir de Descrições de Poços Baseadas em Ontologias. / Luan Fonseca Garcia. – Porto Alegre: Graduação em Ciência da Computação da UFRGS, 2013.

73 f.: il.

Trabalho de Conclusão – Universidade Federal do Rio Grande do Sul. Curso de Ciência da Computação, Porto Alegre, BR–RS, 2013. Orientador: Mara Abel; Co-orientador: Joel Luis Carbonera.

1. Ontologias. 2. Correlação Litológica. 3. Agrupamento de Dados. 4. Alinhamento de Sequências. 5. Inteligência Artificial. I. Abel, Mara. II. Carbonera, Joel Luis. III. Título.

“O poeta é um fingidor.
Finge tão completamente
Que chega a fingir que é dor
A dor que deveras sente.

E os que leem o que escreve,
Na dor lida sentem bem,
Não as duas que ele teve,
Mas só a que eles não têm.

E assim nas calhas de roda
Gira, a entreter a razão,
Esse comboio de corda Que se chama coração.”
— FERNANDO PESSOA

“Como dar vida a uma verdadeira obra de arte
A não ser com a própria vida?”
— MÁRIO QUINTANA

AGRADECIMENTOS

Dedico este trabalho ao meu pai, João Paulo Garcia, que esteve presente no início desta longa jornada, mas infelizmente não poderá presenciar seu final. Sou eternamente grato por tudo que me ensinastes pai! Agradeço à minha mãe, Laura Garcia, minha irmã, Júlia Garcia, e minha noiva, Rosimeri Ribeiro. Não seria possível concluir esta etapa se não houvesse o suporte de minha família. Obrigado a todas vocês, sei que não fui fácil de conviver, por isso que valorizo muito o que fizeram por mim.

Tive a grande felicidade durante esta graduação de conhecer a professora Mara Abel. Mara, não tenho palavras para descrever todo apoio e oportunidade que me destes nestes últimos dois anos. É um grande prazer trabalhar com alguém como você, muito obrigado pela orientação e todos os conselhos que me passaste.

Joel Carbonera, fostes muito mais do que um co-orientador. Aprendi muito com teus conselhos e nossas discussões. Este trabalho não existiria se não fosse por ti! És um grande amigo, muito obrigado por tudo!

Meus grandes amigos e parceiros de muitas disciplinas, Anderson Santos, Alexandre Kreismann, Guilherme Dias, Guilherme Schievelbein, Lucas Freire, Lucas Tomasi e Vinícius Graciolli. Muito obrigado! Tornaram este curso muito mais interessante. Espero poder manter nossa amizade por muito tempo.

Não poderia esquecer também outro grande parceiro durante este curso, meu primo, Lucas Holz. Muitas madrugadas em frente a um computador, programando! Fostes para mim um professor, obrigado!

Gostaria de agradecer também, pela ótima convivência, aos colegas de laboratório e de viagens, do grupo BDI, Sandro Fiorini, Ricardo Werlang, Ricardo Linck e todos outros integrantes do grupo.

Agradeço à Endeeper, por disponibilizar dados e materiais para realização deste trabalho. Agradeço às geólogas Rita e Ana Júlia, por nos ajudarem na tarefa de aquisição de dados e análise dos resultados. Agradeço também à professora Karin Goldberg, pela disposição em ajudar a analisar os resultados que obtivemos e pela aula de geologia ministrada!

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Neste trabalho, propomos uma abordagem automática para o problema da correlação litológica no domínio da Geologia do Petróleo. A correlação litológica visa reconhecer uma mesma dada porção de rocha em diferentes sequências estratigráficas, que geralmente são espacialmente distantes entre si. Tais sequências ocorrem em subsuperfícies, impossibilitando uma avaliação visual direta. Por esta razão, a correlação litológica é utilizada com base em descrições destas sequências, geradas por geólogos a partir de testemunhos, que são cilindros de rocha retirados de poços durante sua perfuração. A análise da continuidade lateral destas porções de rochas, inferida através da correlação, permite estimar a distribuição espacial e o volume de um reservatório, possibilitando a avaliação de sua economicidade.

Em nossa abordagem, adaptamos algoritmos de alinhamento de sequências genéticas, do domínio da Bioinformática, para o domínio da Geologia. Esta adaptação foi realizada com base na suposição de que é possível estabelecer uma equivalência formal entre os problemas de alinhamento e de correlação litológica. Os elementos geológicos que são alinhados nesta abordagem são representados computacionalmente utilizando uma ontologia de domínio que impõe uma estrutura rica e padronizada para as descrições destes objetos, especificando conhecimento sobre litologias, estruturas e texturas de rochas sedimentares, ígneas e metamórficas.

O uso de algoritmos de agrupamento de dados permite realizar abstrações para conjuntos de fácies sedimentares que são similares, viabilizando a realização de comparações de fácies de um modo semelhante às realizadas pelos algoritmos de alinhamento de sequência, quando operam sobre sequências de DNA.

O método desenvolvido foi aplicado sobre dois conjuntos de dados: onze poços de exploração de carvão descritos pela CPRM da Formação Rio Bonito do Rio Grande do Sul e quatro poços de exploração da Formação Santa Luzia no Espírito Santo, pertencentes à Petrobras.

Os resultados mostraram limitações para identificar correlações válidas em termos geológicos. As descrições da CPRM mostraram-se incompletas para caracterizar as unidades em análise. Os poços de Santa Luzia mostraram detalhamento adequado, porém possuem lacunas de testemunhagem justamente na maior parte dos intervalos a serem correlacionados. Essas falhas e as condições geológicas particulares da Formação Santa Luzia dificultaram a obtenção de resultados que permitissem avaliar o algoritmo. Além disso, a partir dos testes foi possível constatar que o fato de não serem considerados pesos de importância distintos para os atributos durante a aplicação de técnicas de agrupamento constitui uma das limitações que versões futuras da abordagem devem superar. Estas limitações serão reavaliadas com novos dados coletados.

Palavras-chave: Ontologias, Correlação Litológica, Agrupamento de Dados, Alinhamento de Sequências, Inteligência Artificial.

1 INTRODUÇÃO

Este trabalho insere-se no contexto do projeto Obaitá, desenvolvido pelo grupo BDI (Grupo de Bancos de Dados Inteligentes da UFRGS). Neste projeto, investigamos abordagens integradas para aquisição, modelagem, representação e raciocínio sobre conhecimento visual. Consideramos conhecimento visual como sendo o conjunto de modelos mentais que suportam o processo de raciocínio sobre informação relacionada ao arranjo espacial e outros aspectos visuais das entidades de domínio (LORENZATTI et al., 2009; CARBONERA, 2012).

Nesta etapa do projeto Obaitá, investigamos abordagens que permitam automatizar a tarefa de correlação litológica, no domínio da Geologia do Petróleo. Na tarefa da correlação litológica, o geólogo busca determinar a continuidade lateral de unidades de rocha a partir da análise de sequências estratigráficas que são espacialmente distantes entre si.

Neste trabalho, parcialmente apresentado em (GARCIA; CARBONERA; ABEL, 2013), propomos uma abordagem automática para a tarefa de correlação litológica. A ideia chave desta proposta é a utilização de versões adaptadas de algoritmos de alinhamento de sequências (SMITH; WATERMAN, 1981), utilizados na Bioinformática, para realizar a correlação litológica. A adaptação destes algoritmos para lidar com o problema em foco neste trabalho foi realizada assumindo uma equivalência formal entre a tarefa de alinhamento de sequências e a tarefa de correlação litológica. É importante notar que, enquanto no alinhamento de sequências busca-se encontrar subsequências de DNA, RNA ou de proteínas semelhantes entre duas ou mais sequências; na correlação de poços busca-se encontrar subsequências de fácies sedimentares semelhantes em dois ou mais poços de petróleo.

As descrições de poços que utilizamos para gerar correlações foram obtidas com o auxílio de uma ontologia de domínio, proposta em (LORENZATTI, 2009) e estendida em (CARBONERA, 2012). Uma ontologia bem fundamentada permite representar explicitamente a semântica dos dados, impondo uma estrutura formal sobre eles que reflete a conceitualização compartilhada por uma determinada comunidade. O uso de ontologias no desenvolvimento de sistemas favorece a integração e o reuso do conhecimento de domínio, e permite realizar o processamento computacional das informações de um modo que reflita como as pessoas concebem os objetos do mundo. Este trabalho é pioneiro em buscar a interpretação automática dos dados gerados a partir desta abordagem.

Um desafio que deve ser enfrentado para realizar a adaptação dos algoritmos de alinhamento para a tarefa em foco neste trabalho está no fato de que algoritmos de alinhamento de DNA realizam comparações triviais entre caracteres. Esta característica constitui um desafio porque os objetos geológicos que são alinhados na correlação litológica são objetos complexos, compostos por diversos atributos, onde uma simples

comparação não é possível. Para contornar este obstáculo, utilizamos uma abordagem de comparação de fácies sedimentares baseada no uso de técnicas de agrupamento de dados.

Agrupamento de dados é a tarefa em que o objetivo é criar grupos (*clusters*) de objetos de maneira que os objetos no mesmo grupo sejam muito similares, enquanto objetos em grupos diferentes sejam muito distintos (GAN; MA; WU, 2007). Utilizando técnicas de agrupamento conseguimos identificar grupos de fácies sedimentares similares. Com isto é possível abstrair fácies em *clusters*, de forma que as comparações de fácies sedimentares sejam realizadas comparando o *cluster* a que pertencem. Isto permite transformar a comparação entre objetos complexos (as fácies sedimentares) em uma comparação entre elementos simples, (os rótulos dos *clusters* a que as fácies pertencem), e possibilita que utilizemos os algoritmos de alinhamento para realizar correlações.

A ontologia utilizada neste trabalho foi desenvolvida seguindo princípios metodológicos da Engenharia de Ontologias, a partir de consultas com especialistas da área, buscando refletir a conceitualização deles sobre os objetos geológicos de interesse nesta abordagem. Esta ontologia oferece um conjunto de diversos atributos e relações que permite representar estes objetos com mais detalhes, em comparação com a representação utilizada em outras abordagens. Este conjunto detalhado de informações descritivas possibilita a comparação de objetos do domínio, impactando na qualidade do resultado das técnicas de agrupamento empregadas.

Realizamos uma implementação desta abordagem proposta. O sistema resultante desta implementação pode ser visto como uma plataforma para pesquisas futuras em abordagens automáticas para correlação litológica, permitindo o teste de novas abordagens. Para realizar testes da abordagem, com o auxílio de geólogos, foram coletadas descrições de poços disponíveis em repositórios públicos ou descritos em relatórios de estudos e projetos e os adequamos à nossa ontologia. Estes dados foram convertidos para os formatos utilizados em nossa implementação, permitindo a realização de testes junto a especialistas.

A organização deste texto é a seguinte. Nos capítulos 2, 3, 4 e 5, revisamos os principais conceitos de ontologias, alinhamento de sequências, agrupamento de dados e geologia, respectivamente. No capítulo 6, realizamos uma revisão dos trabalhos existentes com o objetivo de realizar correlações automáticas. No capítulo 7, apresentamos um fragmento da ontologia de domínio em que foram baseadas as descrições utilizadas neste trabalho. Nos capítulos 8 e 9, apresentamos nossa abordagem e uma proposta de implementação para ela. No capítulo 10, apresentamos e discutimos os resultados obtidos e, no capítulo 11, apresentamos nossas conclusões e possíveis trabalhos futuros.

2 ONTOLOGIAS

A palavra Ontologia tem sua origem no ramo da Filosofia, e podemos entender como sendo um sistema específico de categorias que representa uma determinada visão do mundo. No entanto, na área da IA, este termo tem sido utilizado com outro significado. Guarino (1998) define ontologia como um *artefato de engenharia*, constituído por um vocabulário particular, utilizado para descrever uma realidade, acrescido de proposições relativas ao significado pretendido dos termos deste vocabulário. Apesar de serem diferentes, as visões da filosofia e da IA são relacionadas. Guarino diferencia ambas, mantendo o termo ontologia para referenciar um artefato de engenharia, composto por conceitos e relações, enquanto renomeia a visão filosófica como conceitualização. Sendo assim, duas ontologias podem diferir em sua linguagem, porém compartilhar a mesma conceitualização, ou seja, a mesma visão da realidade.

Uma das definições mais citadas atualmente tem origem em uma união das definições de Gruber (GRUBER, 1993) e de Borst (BORST, 1997), realizada por Studer el at (1998). Studer define uma ontologia como *uma especificação formal e explícita de uma conceitualização compartilhada*. O conhecimento descrito por uma ontologia deve ser explícito e formal, pois desejamos que todo o conhecimento existente do domínio seja processável por computador. Além disto, ela deve refletir uma visão de mundo consensual, expressando o conhecimento de uma comunidade, e não um conhecimento individual.

Ontologias são importantes porque permitem impor uma estrutura formal e homogênea refletindo o conhecimento consensual compartilhado pelos membros de uma comunidade. Isto facilita a construção de sistemas baseados em conhecimento, visto que é possível reutilizar o conhecimento de domínio representado em ontologias. Além disso, favorece a interoperabilidade e integração de sistemas.

Guarino (1998), classifica ontologias em quatro diferentes tipos (Figura 2.1), de acordo com seu nível de generalização. Os tipos são ontologia de topo, ontologia de domínio, ontologia de tarefa e ontologia de aplicação. *Ontologias de Topo*, também chamadas de ontologias de nível superior, descrevem conceitos genéricos, como espaço, tempo, objetos, eventos, entre outros, que são independentes de um domínio particular. *Ontologias de Domínio* e *Ontologias de Tarefa* descrevem um vocabulário relativo a um domínio ou a uma tarefa, através de especializações de ontologias de topo. *Ontologias de Aplicação* descrevem conceitos dependentes tanto de um domínio específico como de uma tarefa específica, que são geralmente especialização de ambas as ontologias relacionadas.

Dentre as ontologias de topo, destaca-se a UFO (Unified Foundational Ontology), uma ontologia baseada em teorias de diversas áreas, como Ontologia Formal, Lógica Filosófica, Filosofia da Linguagem e Psicologia Linguística e Cognitiva (GUIZZARDI, 2005). A UFO foi criada como uma unificação de outras ontologias de fundamentação,

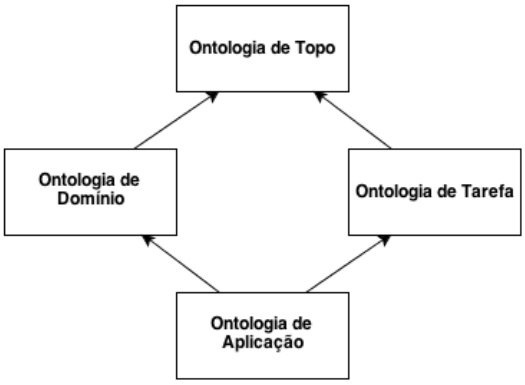

Figura 2.1: Hierarquia de diferentes tipos de ontologias. Traduzida e adaptada de (GUARINO, 1998)

como OntoClean, Dolce e GFO (GUIZZARDI; WAGNER, 2010). Podemos subdividí-la em três partes: UFO A, UFO B e UFO C. A *UFO A* é considerada o núcleo da UFO, e trata de universais e indivíduos duradouros no tempo, chamados de *endurantes*. A *UFO B* é um incremento da parte anterior, e trata de eventos, ou seja, universais e indivíduos que ocorrem durante uma porção de tempo, chamados de *perdurantes*. A *UFO C* acrescenta às duas porções anteriores entidades sociais e intencionais, incluindo entidades linguísticas (GUIZZARDI, 2005). Para o escopo deste trabalho utilizaremos noções do fragmento UFO A, que apresentaremos a seguir. Detalhes do fragmentos UFO B e UFO C podem ser encontrados, respectivamente, em (GUIZZARDI et al., 2013) e (GUIZZARDI et al., 2007).

2.1 UFO - Unified Foundational Ontology

Na UFO, é realizada uma distinção entre Indivíduos (*Individuals*) e Universais (*Universals*). Indivíduos são entidades existentes na realidade e que possuem um critério de identidade único, enquanto Universais são padrões de características, que podem ser instanciados em um número de diferentes indivíduos. Os diversos tipos de universais previstos pela UFO podem ser vistos como metatipos, no sentido de que eles são tipos de tipos de indivíduos. Deste modo, os tipos de indivíduos (representados como classes ou conceitos) em ontologias de domínio, tarefa ou aplicação, podem ser vistos como instâncias destes metatipos estabelecidos pela UFO. Ontologias construídas utilizando a UFO como ontologia de fundamentação comprometem-se com esta meta-conceitualização. Este comprometimento tem como consequência uma diminuição do conjunto de ontologias possíveis, aproximando este conjunto ao conjunto de ontologias pretendidas, de acordo com esta meta-conceitualização.

A UFO A considera apenas Universais Endurantes (*Endurant Universals*), que são aqueles cujos indivíduos *são* no tempo, no sentindo de que estão totalmente presentes, sempre que estão presentes, como *Pessoa*, *Maçã* e *Celular*.

A classificação dos universais existentes na UFO é feita de acordo com as noções de identidade, rigidez, dependência existencial e dependência relacional. O *critério de identidade* é o critério que utilizamos para julgar se duas entidades são a mesma (ou não). Em relação à rigidez, dizemos que uma propriedade é *rígida* se ela é necessária para todas as suas instâncias. Exemplos seriam *Pessoa* e *Animal*; sendo dita *anti-rígida*, caso não seja necessária para qualquer das suas instâncias; ou *semi-rígida*, caso seja necessária para

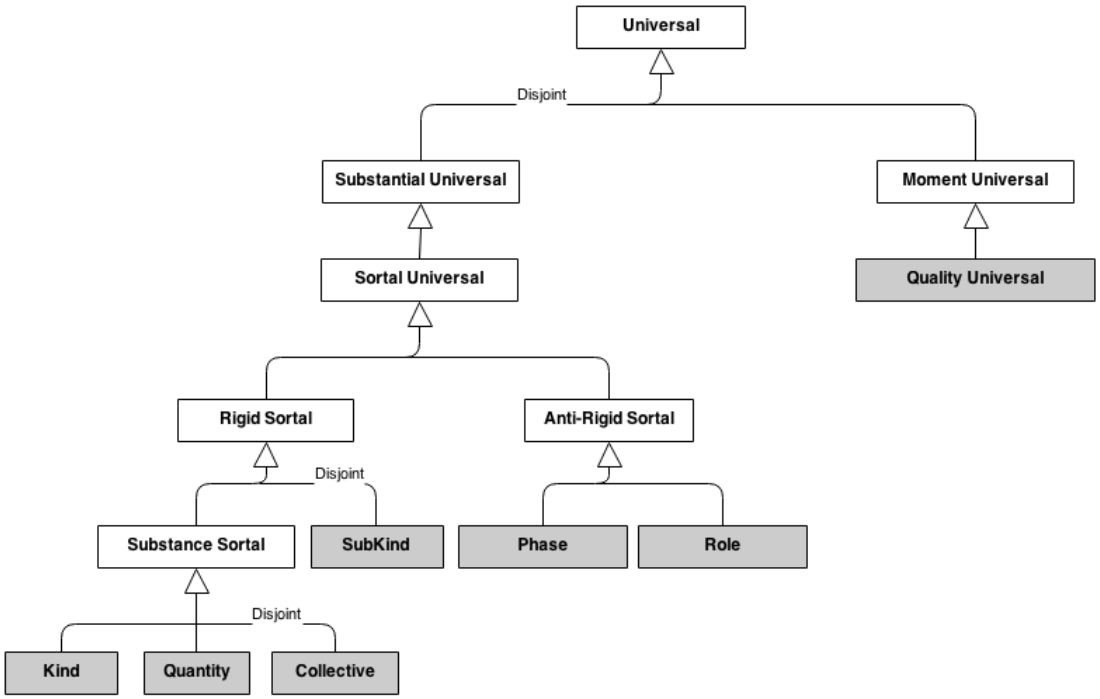

Figura 2.2: Estrutura da UFO A. Adaptada de (GUIZZARDI, 2005)

algumas de suas instâncias, mas contingente (ou acidental) para outras. *Estudante* é um exemplo de uma propriedade anti-rígida, visto que *Estudantes* não são necessariamente *Estudantes*, uma vez que qualquer *Estudante* pode deixar de ser *Estudante* e continuar a existir enquanto *Pessoa*. Já a propriedade *Objeto Sentável* é semi-rígida, pois sua ocorrência é obrigatória para o tipo *Cadeira*, mas acidental para o tipo *Mesa*. Por fim, dizer que uma entidade *x* é *existencialmente dependente* de uma entidade *y*, significa que se a entidade *x* existe, então é necessário que exista a entidade *y*.

Na figura 2.2, estão representados os Universais de Substância (*Substantial Universals*), e os Universais de Momento (*Moment Universals*). Os *Universais de Substância* que são aqueles cujos indivíduos são existencialmente independentes, no sentido de que não são necessários outros indivíduos para existirem. Por exemplo, Pessoa, Livro e Carro podem ser vistos como Universais de Substância. Os *Universais de Momento*, por outro lado, são aqueles indivíduos existencialmente dependentes, ou seja, dependem de outros indivíduos para existir. Como exemplos destes últimos, podemos citar *Cor*, *Peso* e *Altura*.

Dentro dos Universais de Substância, temos a classe dos Universais Sortais (*Sortal Universals*. *Universais Sortais* são aqueles que oferecem um único critério de identidade para todas as suas instâncias.

Universais Sortais podem ser Rígidos (*Rigid Sortal*) ou Anti-Rígidos (*Anti-Rigid Sortal*). Os *Sortais Rígidos* podem ser Sortais de Substância (*Substance Sortal*), e então especializados em Tipo (*Kind*), Quantidade (*Quantity*) e Coletivo (*Collective*), ou podem ser Sub-Tipo (*SubKind*). *Tipos* são complexos funcionais que oferecem seu próprio critério de identidade para suas instâncias, como *Pessoa*, *Macaco*, etc. *Quantidades* representam porções de matéria que apresentam critério de identidade bem definido, como *Água* e *Vinho*. *Coletivos* são coleções de complexos funcionais homogêneas que também possuem critério de identidade próprio, como *Baralho*, *Matilha*, etc. Os *Sub-Tipos* são sortais que não possuem critério de identidade próprios, porém herdam de outros Sortais de Substância. Por exemplo *Homem* e *Mulher*, que herdam seu critério de identidade do

tipo *Pessoa*.

Sortais Anti-Rígidos, por sua vez, são especializados em Fase (*Phase*) ou Papel (*Role*). Fases e Papéis diferem quanto a sua dependência relacional, pois enquanto Fases são relacionalmente independentes, Papéis são relacionalmente dependentes. *Fases* definem partições disjuntas de um conjunto, como por exemplo, *Infância*, *Adolescência* e *Idade Adulta*, de um *Ser Humano*. Um Indivíduo desempenha um *Papel* quando está relacionando-se a uma outra entidade. Por exemplo, uma *Pessoa* desempenha o papel de *Estudante* quando está matriculada em uma *Instituição de Ensino*.

Universais de Momento são especializados por Universais de Qualidade (*Quality Universal*), que são aqueles associados a Estruturas de Qualidade (*Quality Structures*). *Estruturas de Qualidade* podem ser um Domínio de Qualidade (*Quality Domain*) ou uma Dimensão de Qualidade (*Quality Dimension*), composta por diversos domínios. Um exemplo de um domínio de qualidade é a propriedade *Peso*,, cujos indivíduos podem ter valores associados e estes valores são definidos em uma estrutura de qualidade que é isomórfica à parte positiva da linha dos números reais. Um exemplo de uma dimensão de qualidade é a propriedade *Cor*, que pode ser composta pelas dimensões de saturação, brilho e contraste.

Na UFO, existem ainda quatro relações partonômicas definidas. São elas componenteDe (*componentOf*), subQuantidadeDe(*subQuantityOf*), subColeçãoDe (*subCollectionOf*) e membroDe(*memberOf*). A relação *componenteDe* é estabelecida entre universais que são Tipos, como por exemplo, um *Processador* é parte de um *Computador*. A relação *subQuantidadeDe* é estabelecida entre universais que são Quantidades, como por exemplo na relação o *Leite* é parte do *Mingau*. A relação *subColeçãoDe* é estabelecida entre universais que são Coletivos, como por exemplo na relação os *Filhotes* dos *Lobos* são parte da *Matilha*. A relação *membroDe* é estabelecida entre universais que são Tipos ou Coletivos e universais Coletivos, como na relação um *Lobo* é parte da *Matilha*.

Neste trabalho, a ontologia de domínio utilizada (proposta por Lorenzatti(2009) e estendida por Carbonera (2012)) utiliza os metatipos e relações descritos nesta seção. *Fácies Sedimentares* e *Estruturas Sedimentares*, por exemplo, foram classificadas como *Tipos*. Já *Rochas Siliciclásticas*, constituintes das *Fácies Sedimentares*, foram classificadas como *Quantidade*. Esta ontologia de domínio é descrita em maior detalhes no capítulo 7.

3 ALINHAMENTO DE SEQUÊNCIAS BIOLÓGICAS

Alinhamento de sequências no domínio da Bioinformática é o procedimento de comparar duas ou mais sequências biológicas, procurando por uma série de caracteres individuais ou padrões de caracteres que estão na mesma ordem em todas as sequências (MOUNT, 2004). Sejam as sequências $A = \{a_1, a_2, ..., a_n\}$ e $B = \{b_1, b_2, ..., b_m\}$, de tamanhos n e m, respectivamente, em um alinhamento acrescentam-se *gaps* (representados pelo caractere hífen) nas sequências de maneira a maximizar o número de elementos iguais em posições coincidentes. Quando em posições de mesmo índice das duas sequências existem elementos idênticos, temos um *match*. Quando em ambas sequências na mesma posição existem elementos distintos (e não um *gap* em alguma das sequências), temos um *mismatch*. *Uma restrição* importante é que não podem ser alinhados *gaps* de uma sequência com *gaps* de outra sequência. De acordo com Mount (2004), através do alinhamento, é possível obter informações estruturais, funcionais e evolucionárias de organismos biológicos. Além disso, também é possível encontrar ancestrais em comum destes organismos, pois os alinhamentos destas sequências indicam mudanças que podem ter ocorrido entre os dois organismos e seu ancestral.

Segundo Mount (2004), os métodos de alinhamentos podem ser globais ou locais (Figura 3.1). No *alinhamento global* o objetivo é encontrar o maior número de elementos alinhados ao longo de uma sequência inteira. Este tipo de alinhamento é mais apropriado quando as sequências que serão alinhadas possuem tamanho próximo e se espera haver um alto grau de similaridade entre ambas (o que tem como consequência alinhamentos em toda a extensão das sequências). O *alinhamento local* visa encontrar as maiores subsequências comuns entre as sequências. Para este tipo de alinhamento sequências de tamanhos distintos com subsequências similares são mais apropriadas.

Uma das principais abordagens para solução de problemas de alinhamento é a de algoritmos de programação dinâmica. Algoritmos de *programação dinâmica* proveem soluções de complexidade polinomial para uma classe de problemas de otimização que possuem uma substrutura ótima, na qual a solução do problema como um todo pode ser encontrada a partir da solução ótima de subproblemas que se sobrepõem, de maneira que possam ser computados independentemente e memorizados para reuso (MANDOIU; ZELIKOVSKY, 2008). De acordo com Chao et al (2009), este tipo de abordagem foi introduzida no domínio da Bioinformática primeiramente em (NEEDLEMAN; WUNSCH, 1970), com um algoritmo que ficou conhecido como *Algoritmo de Needleman-Wunsch*, utilizado para encontrar alinhamentos globais em duas sequências distintas. Em (1981), Smith modificou o algoritmo de Needleman-Wunsch para encontrar alinhamentos locais nas sequências, criando um algoritmo que ficou conhecido como *Algoritmo de Smith-Waterman*. Este é o algoritmo utilizado para a implementação de nossa proposta, portanto, apresentaremos seu funcionamento em detalhes na próxima seção.

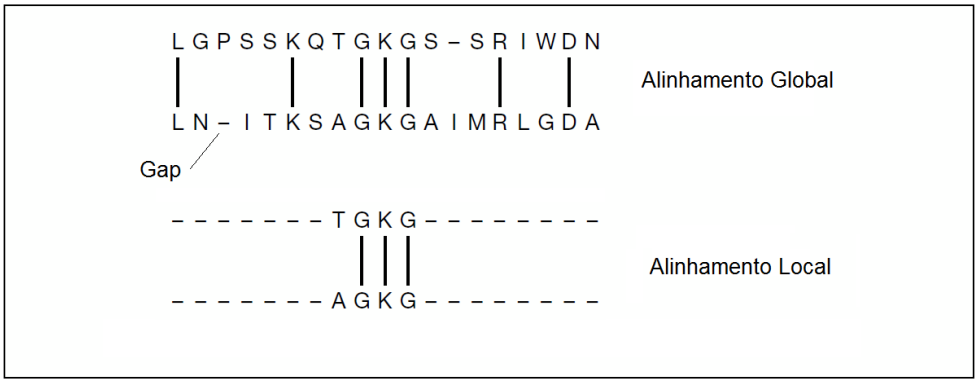

Figura 3.1: Alinhamentos Global e Local. Traduzida e adaptada de (MOUNT, 2004)

3.1 Algoritmo de Smith-Waterman

Dadas duas sequências $A = \{a_1, a_2 ... a_n\}$ e $B = \{b_1, b_2 ... b_m\}$, com n e m sendo o número de elementos das sequências A e B, pesos W_k para *gaps* de tamanho k, e alguma função de similaridade $s(a,b)$ entre os elementos a e b. Esta função tem como retorno valores pré-definidos para *matchs* e *mismatchs*. Para encontrar subsequências com alto valor de similaridade uma matriz H é construída:

$$H_{k0} = H_{0l} = 0 \text{ para } 0 \leq k \leq n \text{ e } 0 \leq l \leq m$$

O restante da matriz é preenchida obedecendo as seguintes regras:

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + s(a_i, b_j) \\ \{H_{i-k,j} - W_k\} \text{ onde k} \geq 1 \\ \{H_{i,j-l} - W_l\} \text{ onde l} \geq 1 \\ 0 \end{cases} \tag{3.1}$$

para $1 \leq i \leq n$ e $1 \leq j \leq m$.

Cada posição H_{ij} da matriz corresponde ao alinhamento com pontuação máxima das sequências terminadas em a_i e b_j. A primeira regra se aplica quando a_i e b_j são *matchs*, a segunda regra quando a_i está no final de um *gap* de tamanho k, a terceira quando b_i está no final de um *gap* de tamanho l e o valor zero é incluído apenas para garantir que não sejam incluídos valores negativos na matriz. A posição da matriz H_{ij} com maior valor representa os elementos finais do alinhamento de maior valor entre as sequências. A partir desta posição uma operação de *traceback* é realizada. Nesta operação, a posição anterior da matriz que resultou no maior valor para a posição atual é a posição a ser percorrida. Este procedimento é realizado recursivamente até que se alcance uma posição de valor zero obtemos o alinhamento.

4 AGRUPAMENTO DE DADOS

Agrupamento (*Clustering*) é um tipo de abordagem de Aprendizado de Máquina Não-Supervisionado (*Unsupervised Learning*) onde o objetivo é criar grupos de objetos, também chamados de *clusters*, de maneira que objetos no mesmo *cluster* sejam muito similares, enquanto objetos em *clusters* diferentes sejam muito distintos (GAN; MA; WU, 2007). *Aprendizado Não-Supervisionado* é um tipo de abordagem de Aprendizado de Máquina que busca aprender padrões em um conjunto de dados em que suas instâncias não são rotuladas (SAMMUT; WEBB, 2011). Segundo Han (2006), técnicas de agrupamento vem sendo aplicadas com sucesso em diversas áreas, como por exemplo o Comércio Virtual, a Biologia e o Processamento de Imagens. No comércio virtual, podemos dividir consumidores em grupos, onde consumidores dentro de um mesmo grupo possuem características e interesses similares, facilitando assim a criação de estratégias de marketing específicas para eles. Na Biologia, podemos utilizar o agrupamento para organizar sequências genéticas similares em grupos, chamados de famílias de genes. No Processamento de Imagens, podemos aplicar algoritmos de agrupamento para dividir regiões de uma imagem digital, visando à detecção de bordas, por exemplo.

Métodos de agrupamento operam sobre *Datasets*, que são conjuntos de objetos em que desejamos encontrar padrões. *Objetos* são entidades, como por exemplo, pacientes de um banco de dados de uma clínica ou estudantes e professores em um banco de dados de uma universidade. Objetos também podem ser chamados de *instâncias*, *amostras* ou *exemplos*. Objetos são descritos como um conjunto de atributos. Um *atributo* é um campo que descreve uma característica de um objeto. Na Tabela 4.1, temos um *dataset* composto por 4 objetos, cada um deles composto por 5 atributos. Em (GAN; MA; WU, 2007), distingue-se os tipos de atributos em contínuos ou discretos (Figura 4.1). *Atributos Contínuos* são aqueles que não possuem um intervalo de valores fechado. *Atributos Discretos* são aqueles que possuem um intervalo de valores definido e fechado. Ainda dentro dos atributos discretos podemos estabelecer uma distinção entre atributos binários e atributos nominais. *Atributos Binários* são aqueles que possuem apenas dois valores possíveis, tal como verdadeiro ou falso, sim ou não e 0 ou 1. *Atributos Nominais* possuem como valores símbolos distintos que servem como nomes ou rótulos.

Segundo (HAN; KAMBER; PEI, 2006), é difícil categorizar métodos de agrupamento devido ao fato de que tais categorias muitas vezes se sobrepõem, pois os métodos provavelmente possuem características de mais de uma destas categorias. Apesar disto, os autores consideram útil organizar as principais técnicas nas seguintes grandes classes: métodos de partição; métodos hierárquicos; métodos baseados em densidade; métodos baseados em grade; e métodos baseados em modelos.

Nos *Métodos de Partição*, dado um conjunto de dados D, seja n o número de objetos, e seja k o número de *clusters* a serem formados, o método particiona os objetos em k

Objetos	Atributos
x1	(A, B, B, D, C)
x2	(D, B, A, D, A)
x3	(C, B, A, E, D)
x4	(B, C, C, C, C)

Tabela 4.1: Tabela representando um *dataset*, onde linhas são objetos e colunas são atributos.

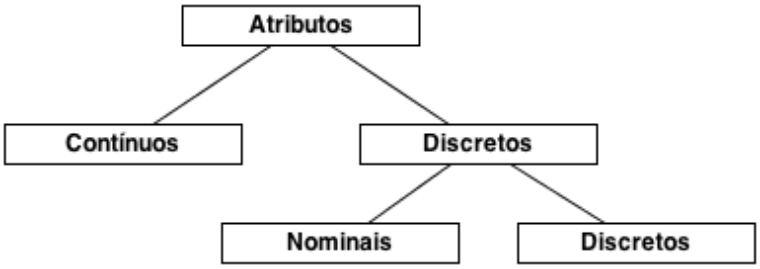

Figura 4.1: Principais tipos de dados dos atributos. Traduzida e adaptada de (GAN; MA; WU, 2007)

partições, para $k \leq n$, onde cada partição representa um *cluster* (HAN; KAMBER; PEI, 2006). Exemplos são o método de *k-means* e *k-medoids*.

Os *Métodos Hierárquicos* criam uma decomposição hierárquica do conjunto de dados. Podem ser classificados como métodos aglomerativos ou divisivos, conforme a decomposição é realizada (HAN; KAMBER; PEI, 2006). No método *divisivo* temos uma abordagem do tipo *top-down*. Todos dados são considerados como pertencentes a um grande *cluster*, e então a cada iteração o método divide o *cluster* em outros *clusters* menores, até que um critério de parada seja satisfeito ou que os *clusters* tenham tamanho 1. Já no método *aglomerativo* a abordagem é do tipo *bottom-up*. Cada objeto do conjunto de dados é considerado um *cluster*, então o método a cada iteração combina os *clusters* até que um critério seja satisfeito ou apenas um *cluster* exista. Exemplos de métodos hierárquicos são o método *AGNES* (aglomerativo) e o método *DIANA* (divisivo).

Métodos Baseados em Densidade identificam *clusters* distintos nos dados, baseados na noção de que um *cluster* num espaço de dados é uma região contígua de alta densidade de objetos, separados de outros *clusters* como tal, por regiões de baixa densidade de objetos. Os dados situados nas regiões de baixa densidade são considerados ruído (SAMMUT; WEBB, 2011). Segundo Han et al (2006), tais métodos são utilizados para encontrar *clusters* de formas diferentes das usuais, como por exemplo, formas ovais e formas em "S", como na figura 4.2. Exemplos de métodos baseados em densidade são o *DBSCAN* e o *OPTICS*.

Métodos Baseados em Grade quantizam o espaço do objeto em um número finito de células que formam uma estrutura de uma grade. A principal vantagem deste tipo de método é que todas as suas operações são realizadas na grade, tornando assim seu tempo de processamento rápido, pois depende apenas do número de células em cada dimensão do espaço quantizado, e não do número de objetos de dados (HAN; KAMBER; PEI, 2006). Tal técnica pode ser aplicada a diversos problemas de mineração de dados, porém em agrupamento geralmente é integrada com métodos hierárquicos ou baseados em densidade. Exemplos são o método *STING* e o método *CLIQUE*.

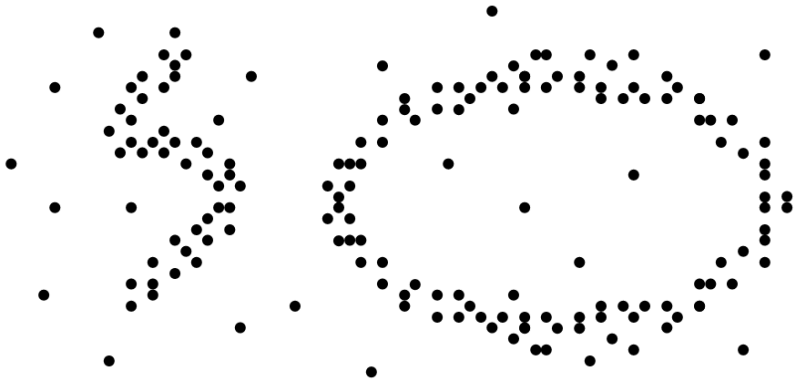

Figura 4.2: Clusters de forma em "S"e de forma oval. Retirada de (HAN; KAMBER; PEI, 2006)

Métodos Baseados em Modelos geram modelos hipotéticos para cada um dos *cluster* e encontram a melhor adequação dos dados para tal modelo. Um algoritmo baseado em modelo pode encontrar *clusters* construindo uma função de densidade que reflete a distribuição espacial dos dados. Além disso, pode levar a uma maneira de determinar automaticamente o número de *clusters* baseado em padrões estatísticos, levando em consideração ruídos nos dados e assim produzindo métodos muito robustos (HAN; KAMBER; PEI, 2006). Exemplos de métodos baseados em modelos são o *EM* e o *COBWEB*.

Na implementação de nossa abordagem, utilizamos o algoritmo Expectation Maximization (DEMPSTER; LAIRD; RUBIN, 1977), pois é um algoritmo simples, onde não é necessário conhecimento a priori do número de clusters e apresentou resultados satisfatórios no agrupamento de nossos dados, segundo especialistas. Seu funcionamento será explicado na seção a seguir.

4.1 Algoritmo Expectation Maximization

O algoritmo EM é um algoritmo baseado em modelos, utilizado para encontrar a estimativa de maior verossimilhança (*likelihood*) dos parâmetros em modelos estocásticos, onde o modelo depende de variáveis não observadas (SAMMUT; WEBB, 2011). Ao invés de termos um espaço de dimensões, onde cada objeto de dado pertence somente a um *cluster*, como nos métodos de partição, podemos representar cada *cluster* como sendo uma distribuição probabilística parametrizada. Cada distribuição pode ser chamada de componente, enquanto o conjunto total destas distribuições é chamado de modelo de mistura de densidades. Desta maneira, podemos agrupar os dados utilizando um número finito de k distribuições, onde cada uma representa um *cluster*. Segundo (HAN; KAMBER; PEI, 2006), o problema é estimar os parâmetros das distribuições probabilísticas de maneira que melhor organize os dados. O EM é uma maneira iterativa de estimar tais parâmetros. O algoritmo funciona como se segue:

1. Selecionar randomicamente k objetos para representarem o centro dos *clusters*, além de escolher valores iniciais para os demais parâmetros das distribuições, também randomicamente.

2. Iterativamente refinar os parâmetros (*clusters*) baseado em duas etapas:

(a) **Etapa de Expectativa**: Atribuir cada objeto de dados x_i a um *cluster* C_k, com i sendo o total de objetos e k sendo o total de *clusters*, com a seguinte probabilidade:

$$P(x_i \in C_k) = p(C_k|x_i) = \frac{p(C_k)p(x_i|C_k)}{p(x_i)} \tag{4.1}$$

onde $p(x_i|C_k) = N(m_k, E_k(x_i))$ segue a distribuição Gaussiana sobre a média m_k, com expectativa E_k. Esta etapa calcula a probabilidade de cada objeto x_i pertencer a cada *cluster* C_k.

(b) **Etapa de Maximização**: Após a etapa anterior, usar os novos parâmetros estimados para refinar os modelos de *clusters*, como abaixo:

$$m_k = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i P(x_i \in C_k)}{\sum_{j=1}^{k} P(x_i \in C_j)} \tag{4.2}$$

onde n é o número total de distribuições e j o índice de um outro *cluster*. Nesta segunda etapa, o resultado é a maximização da medida de *verossimilhança* das distribuições.

5 GEOLOGIA DO PETRÓLEO

O contexto deste trabalho está inserido no domínio da Estratigrafia Sedimentar. *Estratigrafia Sedimentar* é o estudo das relações temporais e espaciais entre corpos de rochas sedimentares (GLUYAS; SWARBRICK, 2009). Segundo Hyne (2001), *Rochas Sedimentares* são aquelas compostas por sedimentos de três tipos: clásticos, orgânicos e cristalinos. Sedimentos clásticos são partículas provenientes do rompimento de outras rochas e que foram transportadas e depositadas. Sedimentos orgânicos são os formados biologicamente, como conchas do mar. Sedimentos cristalinos são formados pela precipitação de sal ou de água. À medida que estes sedimentos são depositados na superfície se tornam sólidos, e portanto, rochas sedimentares.

A descrição de testemunhos de poços de exploração de petróleo ou mineração é uma tarefa tradicionalmente manual e subjetiva, gerando textos com um vocabulário não formal que são complementados por desenhos com formato livre, como exemplificado na Figura 5.1. Dependendo da sua escola de formação e seu interesse de pesquisa, o geólogo vai impor à descrição um foco, formato e rigor descritivo distinto. Como resultado, um dos dados mais importantes para a caracterização de reservatórios é de difícil integração e só é passível de processamento humano. Por isso, a correlação litológica para exploração de petróleo é principalmente feita com dados de logs de petrofísica obtidos ao longo dos poços. A estratégia é limitada pelo fato de que logs geofísicos refletem apenas os tipos litológicos e a granulometria da rocha, não sendo capazes de diferenciar os aspectos texturais e as estruturas de deposição que são informações essenciais para a interpretação estratigráfica.

Com o desenvolvimento dos estudos sobre ontologias, tornou-se possível o desenvolvimento de modelos formais capazes de oferecer suporte para a definição de formas de representação de conhecimento fortemente visuais. Em especial, os trabalhos de Lorenzatti (2009) e Carbonera (2012) ampliaram o conhecimento sobre a descrição de informações visuais em Geologia Sedimentar, que levaram a construção de sistemas de descrição baseados em ontologias de domínio que capturam dados de litologia, texturas, estruturas sedimentares e fósseis e os disponibilizam em um formato processável por computador. Esses formatos de descrição e armazenamento de dados, que serão descritos no Capítulo 8, permitiram a investigação de estratégias de mineração de dados para interpretação geológica tratadas neste trabalho.

Estamos interessados em uma abordagem automática para extrair correlações litológicas a partir de descrições de poços de petróleo. Estas descrições contemplam rochas presentes na subsuperfície da Terra, onde a observação direta não é possível. A correlação busca determinar se os poços cortam camadas de rocha que foram depositadas no mesmo tempo geológico e pelo mesmo evento de deposição. A correlação litológica busca determinar a correlação considerando apenas a similaridade entre pacotes de rochas

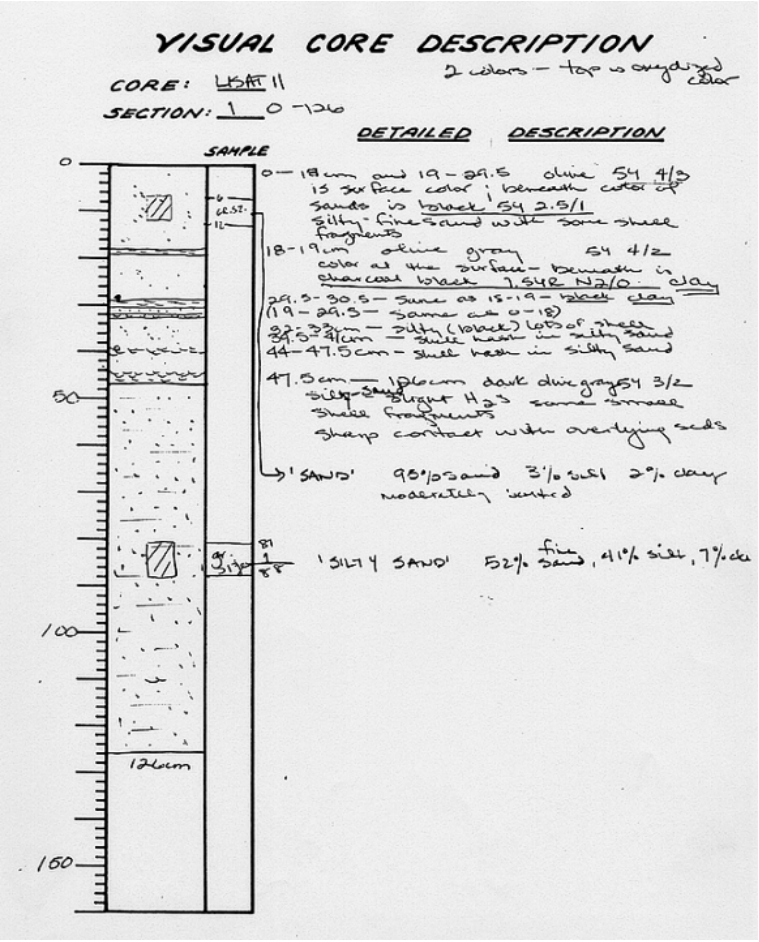

Figura 5.1: Descrição dos atributos visuais de um testemunho de poço de exploração. Extraído de (USGS, 2013)

.

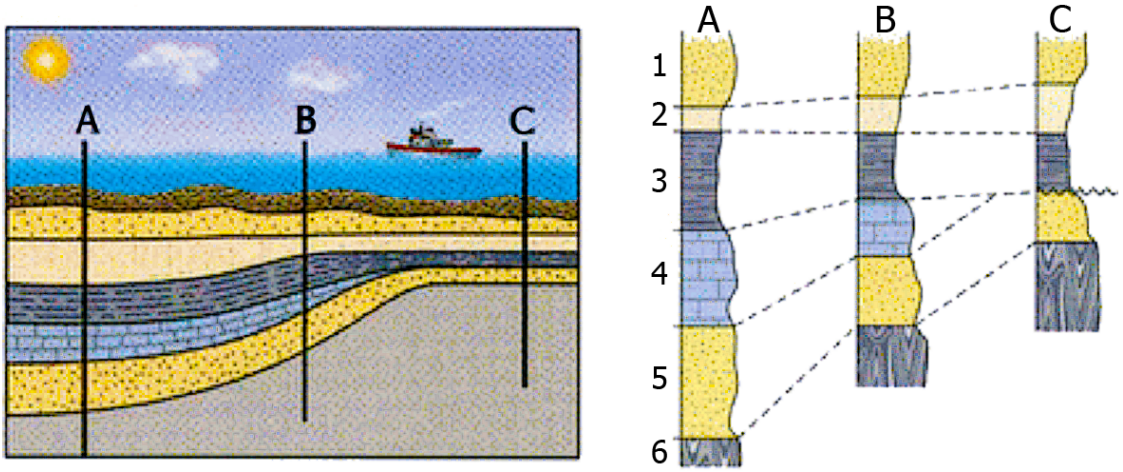

Figura 5.2: Exemplo de correlação litológica. Extraída de (PARSONS, 2013).

descritos a partir de testemunhos e se essas têm continuidade lateral, sem considerar as variações que podem acontecer dentro da deposição em um mesmo evento. A correlação é importante porque permite o entendimento do modelo e do volume de um reservatório de petróleo, o que subsidia a análise de sua economicidade. Um exemplo de correlação pode ser observado na figura 5.2.

Testemunhos de rocha são corpos de rocha cilíndricos (Figura 5.3, retirados de poços de exploração. De acordo com Gluyas (2009), testemunhos são fundamentais para o estudo de um reservatório, pois são as fontes de informações mais precisas sobre o reservatório, além disso, muitas das informações obtidas através do testemunho não são possíveis de serem obtidas através de outras maneiras.

As descrições realizadas pela análise de testemunhos são, na realidade discretizações de corpos de rocha. Uma *Fácies Sedimentar* é uma porção de rocha com características específicas que idealmente deve ser uma rocha distintiva, formada sob certas condições de sedimentação, refletindo um certo processo deposicional ou químico, em um conjunto de condições ou ambiente específico (READING, 2009). O conceito de fácies é usado desde que geólogos, engenheiros e mineiros perceberam que características visuais de unidades de rochas particulares eram úteis para a correlação e predição da ocorrência de carvão, petróleo e minerais. Seu uso moderno foi proposto por (GRESSLY, 1841). Na figura 5.4 é possível visualizar um corpo de rocha com duas fácies distintas. A seguir, detalharemos os principais atributos de fácies sedimentares que serão considerados neste trabalho.

Litologia: Trata-se da composição química e mineral da rocha e suas características texturais. É muito importante para a interpretação de como uma rocha sedimentar foi formada (HYNE, 2001).

Estruturas Sedimentares: São todos os tipos de superfícies, acamadamentos ou estratificações gerados ao longo da deposição sedimentar (PRESS et al., 2004).

Granulometria: É a classificação baseada no diâmetro médio das partículas das rochas sedimentares. O tamanho das partículas indica a energia da corrente de água ou vento que as transportou e auxilia na compreensão do ambiente de deposição em que a rocha foi formada.

Figura 5.3: Testemunhos de rocha. Extraída de (LORENZATTI et al., 2009)

Arredondamento: É a medida do grau de angularidade dos cantos das partículas, que indica a duração do transporte da partícula, ao longo do qual ela foi desgastada, e a distância da área fonte.

Esfericidade: Refere-se ao quanto uma partícula assemelha-se a uma esfera. Na figura 5.5 é possível ver a diferença entre *Arredondamento* e *Esfericidade*.

Cor: Indica a composição dos minérios da rocha. Sozinha não é suficiente para diferenciar fácies (SUGUIO, 2003).

Seleção: Refere-se à segregação dos sedimentos conforme seu tamanho. Uma rocha com partículas de tamanho uniforme é bem selecionada, indicando normalmente um processo mais longo de deposição, enquanto uma rocha que possua partículas de tamanhos variados é dita pobremente selecionada e resultado de um processo de deposição rápido (PRESS et al., 2004).

No próximo capítulo, apresentaremos algumas das abordagens para correlação automática semelhantes à apresentada neste trabalho. Ao final do capítulo, realizaremos uma comparação entre o nosso trabalho e os trabalhos apresentados.

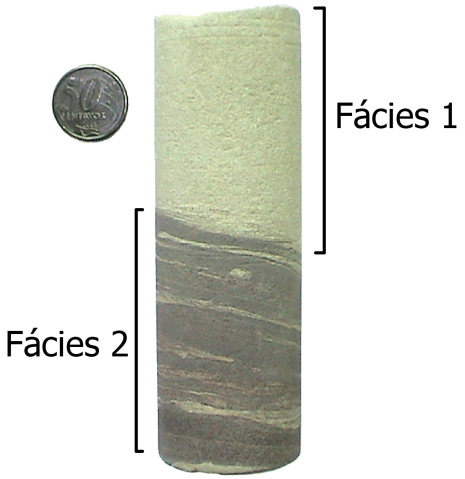

Figura 5.4: Trecho de testemunho de poço, com duas fácies distintas. Adaptada de (LORENZATTI et al., 2009)

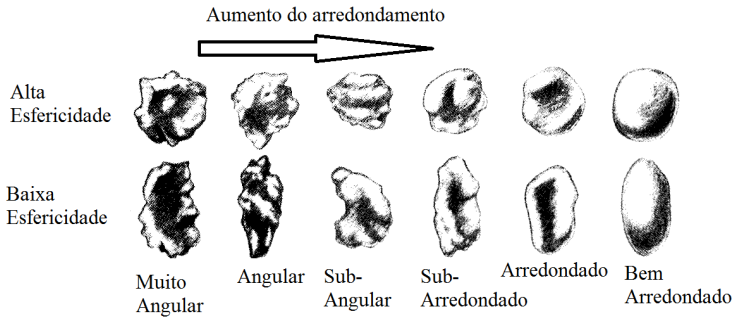

Figura 5.5: Escala de Arredondamento e Esfericidade. Adaptada de (SELLEY, 1998)

6 ABORDAGENS PARA CORRELAÇÃO AUTOMÁTICA EM GEOLOGIA

Devido à semelhança entre as tarefas de alinhamento de sequências e correlação, diversos trabalhos para realizar correlações automáticas foram propostos utilizando algoritmos de alinhamento. Um dos algoritmos de alinhamento de sequência mais utilizados para correlação é o proposto em (SMITH; WATERMAN, 1981). Um dos motivos que levaram à utilização deste algoritmo é o fato de que ele prevê a possibilidade de existirem *gaps*, nas sequências, o que pode representar, no alinhamento litológico, a inexistência de continuidade lateral de fácies sedimentares. Além disso, este algoritmo realiza alinhamentos locais; característica que se ajusta às necessidades deste problema, visto que os poços costumam ter tamanhos diferentes (em termos de número de fácies) e é esperado que existam apenas alinhamentos em porções locais dos poços e não em sua totalidade.

Em Howell (1983), é proposta uma adaptação do algoritmo de Smith-Waterman para realizar correlações. Howell estende o algoritmo de alinhamento de forma que um elemento de uma sequência possa ser correlacionado a mais de um elemento em outra sequência, desta forma, temos um *match* de 1 para n elementos. Isto reflete o fato que porções de rocha podem sofrer um afinamento, de forma que seja visível apenas uma porção em um outro poço. Os dados de entrada são descrições de poços a partir de testemunhos. Estas descrições contemplam apenas a litologia e a granulometria destas unidades. Para a comparação de unidades de rochas dos poços, é calculada uma distância entre as unidades de rocha, com valores definidos em uma matriz de custos granulométrica. Além disso, litologias e espessura das unidades também são utilizadas na comparação.

Waterman (1987) dá continuidade ao trabalho de Howell. O algoritmo é estendido de forma que elementos em ambas sequências possam ser alinhados a mais de um elemento. Deste modo, o algoritmo é capaz de estabelecer um *match* de n para m elementos. Os dados de entrada e a forma de representá-los são os mesmos do trabalho de Howell. As comparações entre elementos também são realizadas calculando a diferença granulométrica, litológica e de espessura das unidades.

Em (WU; NYLAND, 1987), também é utilizado um algoritmo de alinhamento para realizar correlações. A diferença entre os dois trabalhos anteriores é a proveniência dos dados. Enquanto nos trabalhos de Howell (1983) e Waterman (1987), os dados são descrições de testemunhos, Wu e Nyland convertem dados provenientes de logs petrofísicos, como logs de gama, porosidade e resistividade, em unidades baseando-se na granulometria e classe genética das rochas. Logs petrofísicos são logs numéricos que refletem uma medida de alguma propriedade física das rochas capturada através de sondas

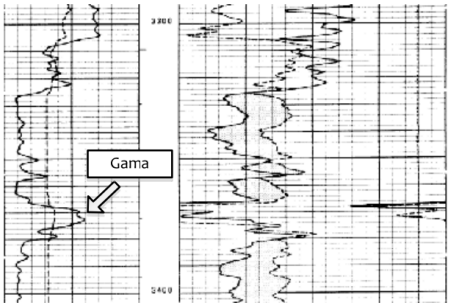

Figura 6.1: Exemplo de log petrofísico. Em destaque a medida de raio gama. Adaptado de (FIORINI, 2009)

.

que são inseridas dentro de poços (Figura 6.1). Esta conversão é necessária porque os dados contidos neste tipo de logs são puramente numéricos. Um problema, apontado em (GRIFFITHS; BAKKE, 1990), é que tal conversão introduz mais incertezas do que necessário.

Griffiths (1990) também propõe que a correlação seja realizada adaptando-se algoritmos de alinhamento. Assim como no trabalho de Wu e Nyland (1987), os dados utilizados para correlação são provenientes de logs de petrofísica. A diferença, entretanto, é que Griffiths não converte os dados de logs, mas sim utiliza diretamente estes dados numéricos para realizar as comparações entre as unidades.

Na Tabela 6.1, apresentamos uma comparação das diferentes abordagens apresentadas e ainda a nossa abordagem. A comparação é feita entre o tipo de dado geológico utilizado, a comparação entre os elementos da sequência realizada e o tipo de *match* realizado por cada abordagem.

Com a análise da comparação das abordagens apresentadas notamos uma carência de fundamentação para a escolha e representação das informações geológicas que são utilizadas para realizar o alinhamento. Além disso, também notamos que as comparações costumam ser realizadas com base em métricas pré-definidas, com base em poucos atributos. A abordagem proposta neste trabalho busca contribuir principalmente na escolha e representação de informações geológicas consideradas e na abordagem de realizar comparações.

Enquanto os trabalhos de Howell (1983) e Waterman (1987) utilizam descrições de testemunhos, sem nenhuma restrição ontológica, e os trabalhos de Wu e Nyland(1987) e de Griffiths (1990) utilizam dados de logs petrofísicos, ou seja, numéricos e não qualitativos, em nosso trabalho, utilizamos descrições de testemunhos com uma formalização imposta por uma ontologia de domínio bem fundamentada. Esta ontologia prevê uma grande quantidade de atributos e valores utilizados pelos geólogos para descrever fácies sedimentares, os quais foram capturados com o auxílio de especialistas, refletindo um conhecimento consensual na área. As descrições resultantes aproximam a forma como os geólogos concebem esses objetos.

Esta grande variedade de atributos definidos pela ontologia auxiliam a desenvolver abordagens de comparação desses objetos geológicos que se aproximem da forma como os geólogos os comparam. Particularmente, em nossa abordagem baseada em agrupamento, os algoritmos têm à disposição um conjunto mais amplo de atributos caracterizadores dos objetos, em comparação com os utilizados pelas abordagens anteriores. Além disso, é importante salientar que os conceitos, relações, atributos e valores de atributos

Trabalho	Tipo de Dados	Comparação de Unidades Litológicas	Match
Howell	Descrições de testemunhos à mão livre	Comparação de granulometria e espessura das unidades	1 para *n* elementos
Waterman	Descrições de testemunhos à mão livre	Comparação de granulometria e espessura das unidades	*n* elementos para *m* elementos
Wu	Logs petrofísicos de poços	Comparação granulométrica e de gênese	1 para 1
Griffiths	Logs petrofísicos de poços	Comparação de logs	1 para 1
Este Trabalho	Descrições de testemunhos baseadas em ontologias	Comparação por agrupamento de dados considerando diversos atributos da ontologia	1 para 1

Tabela 6.1: Tabela de comparação de abordagens computacionais para correlação litológica.

especificados em nossa ontologia buscam representar o conhecimento compartilhado na comunidade. As abordagens anteriores não demonstraram esta preocupação.

É importante notar que existem ainda trabalhos que não se baseiam em algoritmos de alinhamento para realizar correlações, como em (CHEN et al., 1997), (LIM; KANG; KIM, 1998) e (OLEA, 2004), porém seu funcionamento foge ao escopo deste trabalho.

7 ONTOLOGIA DE REPRESENTAÇÃO DE CONHECIMENTO VISUAL PARA A ESTRATIGRAFIA SEDIMENTAR

Os poços de exploração utilizados neste trabalho foram descritos com o auxílio de uma ontologia de domínio para a Estratigrafia Sedimentar, desenvolvida por Lorenzatti em (2009) e estendida por Carbonera em (2012). Os objetos selecionados para modelagem são aqueles que possuem expressão visual que pode ser descrita pelos geólogos. São portanto, sortais de substância de acordo com a classificação da UFO (GUIZZARDI, 2005). Essa restrição responde ao rigor metodológico da abordagem, uma vez que a ontologia de descrição não deveria contemplar conceitos resultantes de alguma interpretação subjetiva do geólogo. A ontologia foi desenvolvida utilizando a ontologia de fundamentação UFO, descrita na seção 2.1, aliada a dois metaconstrutos propostos por Lorenzatti. Os metaconstrutos se fazem necessários porque existem conceitos em que seu significado só pode ser expresso completamente através de representações simbólicas e pictóricas. Ambos foram classificados quanto às propriedades de Identidade, Rigidez e Dependência Existencial, apresentadas no capítulo 2, além da propriedade Unicidade, que será apresentada neste capítulo.

A seguir, apresentaremos os metaconstrutos propostos, *Pictorial Concept* e *Pictorial Attribute*, e um fragmento da ontologia, composto por seus *Universais Endurantes*, *Universais de Qualidade* e *Metaconstrutos Visuais*. Para um detalhamento da ontologia, o leitor deve consultar (LORENZATTI, 2009).

7.1 Metaconstrutos para Representação de Conhecimento Visual

O primeiro metaconstruto é chamado de *Pictorial Concept*. O seu objetivo é representar conceitos que são tipos visuais, que não são atributos ou qualidades de outros conceitos. Os conceitos representados por este metaconstruto ou fornecem Critério de Identidade, ou carregam Critérios de Identidade fornecidos por seus supertipos, possuem rigidez ontológica. Com relação à ontologia de topo UFO, os conceitos representados pertencem ao seguinte conjunto: Tipo (Kind), Quantidade (Quantity), Coletivo (Collective) e Sub-Tipo(SubKind). Na figura 7.1, extraída de (LORENZATTI, 2009), é apresentado um exemplo de aplicação para o metaconstruto para representar uma estrutura sedimentar.

O segundo metaconstruto apresentado é chamado de *Pictorial Attribute*, e tem o objetivo de representar valores de dimensões de qualidade que são associadas a conceitos inerentes a outros conceitos. Os atributos representados por este metaconstruto ou possuem, ou carregam Critérios de Identidade, possuem rigidez ontológica, não possuem critério de unicidade e sempre possuem dependência existencial. Um Pictorial Attribute

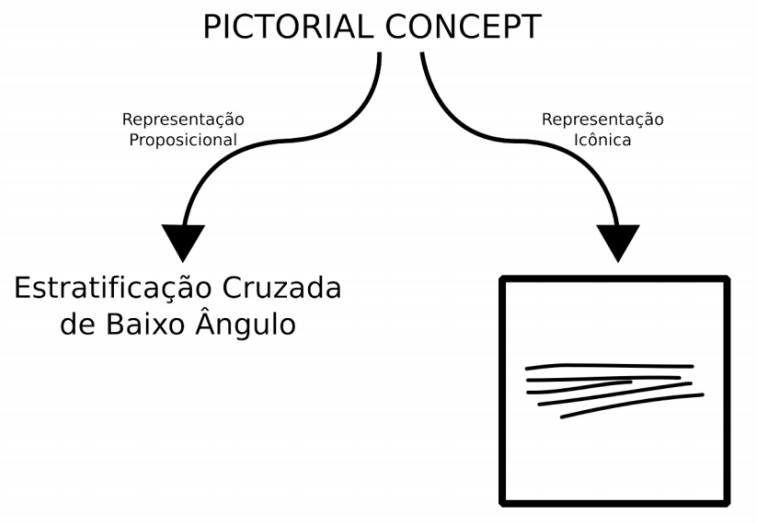

Figura 7.1: Aplicação de um Pictorial Concept para representar uma estrutura sedimentar.

representa valores de Dimensões de Qualidade (Quality Domains) da UFO. Na figura 7.2, extraída de (LORENZATTI, 2009), é apresentado um exemplo de um Pictorial Attribute para representar um valor de um atributo de angularidade.

7.2 Uma Ontologia para Estratigrafia Sedimentar

Como já dito anteriormente, a ontologia foi representada utilizando metaconstrutos da UFO e metaconstrutos propostos por Lorenzatti. Os conceitos modelados foram analisados em relação às metapropriedades, apresentadas no capítulo 2, Identidade, Rigidez e Dependência Relaciona, além da propriedade Unicidade. Identidade é o critério que utilizamos para julgar se duas entidades são a mesma; Rigidez indica se a propriedade é ou não necessária para todas as suas instâncias; Unicidade permite identificar partes e limites de um indivíduo; Dependência Existencial identifica se todo o indivíduo de um universal precisa de algum outro indivíduo para existir.

Dividimos esta seção em três subseções. Na subseção 7.2.1, apresentamos os Universais Endurantes, na sequência, nas subseções 7.2.2 e 7.2.3, apresentamos, respectivamente, os conceitos modelados como Universais de Qualidade e suas expressões visuais, modeladas com o auxílio dos Metaconstrutos Visuais.

7.2.1 Universais Endurantes

Universais Endurantes são aqueles cujos indivíduos *são* no tempo, no sentindo de que estão totalmente presentes, sempre que estão presentes, ou seja, sua identidade não varia ao longo do tempo. Os principais conceitos modelados encontram-se na tabela 7.1. Uma propriedade pode fornecer seu próprio princípio de identidade (+I), pode herdar este princípio (+O) ou pode não possuir este princípio (-I). Em relação à unicidade, uma propriedade pode ser unitária e fornecer seu próprio princípio (+U), pode ser unitária e não fornecer o princípio de unicidade (-U), ou pode não necessariamente ser unitária e não carregar seu princípio, possuindo assim antiunicidade (~U). Em relação à rigidez, uma propriedade pode ser rígida (+R), semirígida (-R) ou antirígida (~R). E finalmente, em relação à dependência relacional, uma propriedade pode ser relacionalmente dependente (+D) ou não relacionalmente dependente (-D).

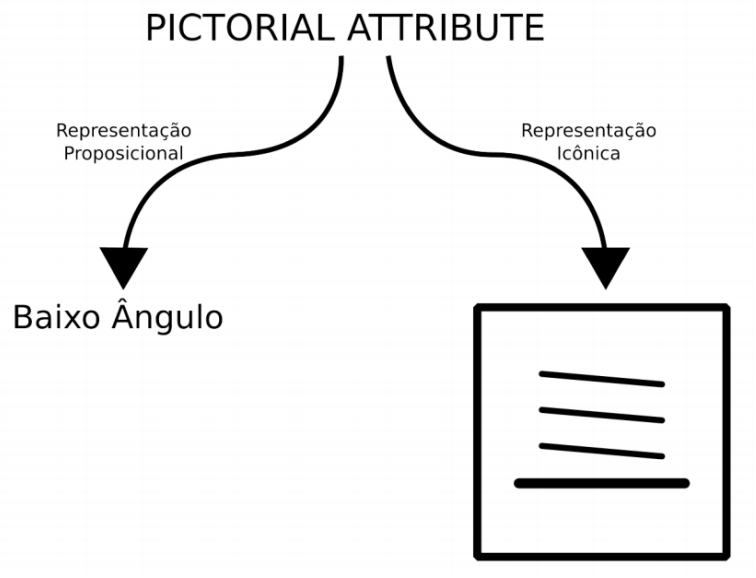

Figura 7.2: Exemplo de aplicação do metaconstruto para representar um valor possível do atributo de angularidade.

Conceito	Identidade	Rigidez	Unicidade	Dependência Existencial
Rocha Siliciclástica	+O	+R	~U	-D
Fácies Sedimentar	+O	+R	+U	-D
Estrutura Sedimentar	+O	+R	+U	-D

Tabela 7.1: Alguns conceitos presentes na ontologia.

Fácies Sedimentares são classificadas na UFO como *Tipo*, e são constituídas por Rocha Siliciclástica e Estruturas Sedimentares. Fácies são universais que oferecem seu próprio princípio de identidade, seu critério de unicidade e são rígidos, no sentido de que uma fácies não poderá deixar de ser uma fácies, ou deixará de existir. São caracterizadas por universais de qualidade como *Granulometria*, *Litologia*, *Cor*, *Seleção*, *Arredondamento* e *Esfericidade*. Estes universais serão detalhados na próxima seção.

Estruturas Sedimentares também são classificadas na UFO como *Tipo*, e o padrão visual formado pelo arranjo espacial dos grãos de sedimentos é sua principal característica.

7.2.2 Universais de Qualidade

Universais de Qualidade são aqueles associados a Estruturas de Qualidade. Eles representam qualidades de outros universais, e seus indivíduos são existencialmente dependentes de indivíduos dos universais que eles caracterizam. *Estruturas de Qualidade* podem ser um Domínio de Qualidade ou uma Dimensão de Qualidade, esta, por sua vez, composta por diversos domínios.

Os conceitos apresentados nesta seção são classificados como Universais de Qualidade porque representam atributos e qualidades de fácies sedimentares, apresentadas na seção anterior, e podem ser visualizados na tabela 7.2.

Universal de Qualidade	Descrição
Arredondamento	É a medida do grau de angularidade dos cantos das partículas. Seus valores possíveis são: *Muito Angular*, *Angular*, *Sub-Angular*, *Sub-Arredondado*, *Arredondado* e *Bem Arredondado*.
Cor	Indica a composição dos minérios da rocha. Possui 119 valores, de acordo com a tabela padrão para rochas NBS/ISCC RC.
Esfericidade	Refere-se ao quanto uma partícula assemelha-se a uma esfera. Seu espaço de valores é: *Alta*, *Média* e *Baixa*
Granulometria	É a classificação baseada no diâmetro médio das partículas das rochas sedimentares. Seu espaço de valores é: *Silte*, *Argila*, *Areia Muito Fina*, *Areia Fina*, *Areia Média*, *Areia Grossa*, *Areia Muito Grossa*, *Cascalho*, *Grânulo*, *Seixo*, *Bloco* e *Matacão*.
Litologia	É a composição química e mineral da rocha e suas características texturais.
Seleção	Refere-se à segregação dos sedimentos conforme seu tamanho. Possui como valores *Muito Bem Selecionado*, *Bem Selecionado*, *Moderadamente Selecionado*, *Mal Selecionado* e *Muito Mal Selecionado*.

Tabela 7.2: Universais de Qualidade intrínsecos a Fácies Sedimentares.

7.2.3 Metaconstrutos Visuais

Tendo classificado os conceitos com relação aos metaconstrutos da UFO, agora é possível classificá-los com os metaconstrutos propostos por Lorenzzati.

O conceito de Estrutura Sedimentar é considerado um tipo visual, e portanto, classificado como um *Pictorial Concept*. Na figura 7.3, podemos observar diferentes tipos de Estruturas Sedimentares.

Arredondamento, Esfericidade e Seleção são conceitos relativos ao conceito de Fácies Sedimentares, sendo assim classificados como *Pictorial Attributes*. A representação visual de suas respectivas dimensões de qualidade pode ser vista nas figuras 7.4, 7.5 e 7.6.

Após a revisão dos conceitos presentes neste trabalho e a apresentação de um fragmento da ontologia em que as descrições aqui utilizadas foram baseadas, nos próximos capítulos apresentaremos nossa abordagem para realizar correlações litológicas (Capítulo 8), uma proposta de implementação para esta abordagem (Capítulo 9) e os resultados obtidos através de experimentos realizados com esta implementação (Capítulo 10).

Figura 7.3: Diferentes representações de estruturas sedimentares. Extraído de (LORENZATTI, 2009)

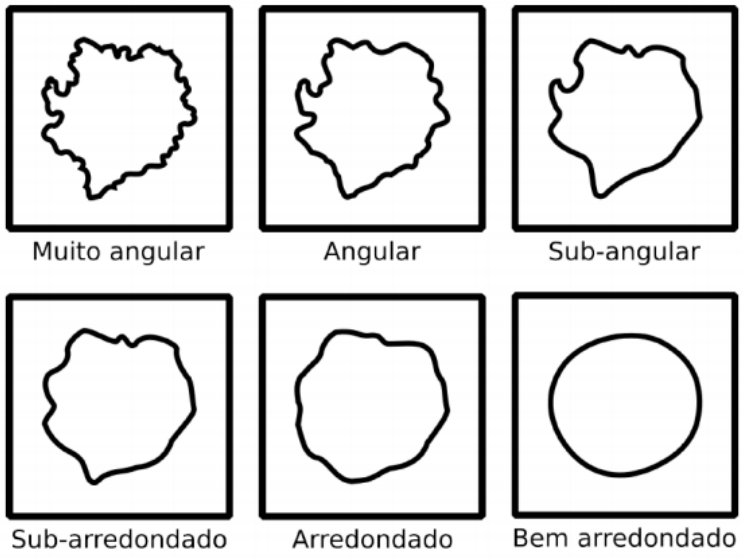

Figura 7.4: Dimensão de Qualidade do Universal de Qualidade Arredondamento. Extraído de (LORENZATTI, 2009)

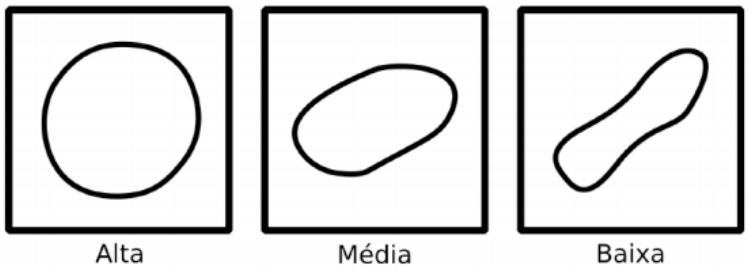

Figura 7.5: Dimensão de Qualidade do Universal de Qualidade Esfericidade. Extraído de (LORENZATTI, 2009)

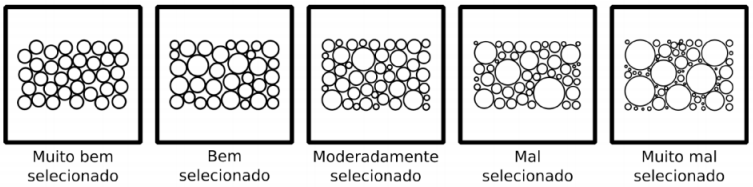

Figura 7.6: Dimensão de Qualidade do Universal de Qualidade Seleção. Extraído de (LORENZATTI, 2009)

8 ABORDAGEM PROPOSTA

Nossa proposta consiste em uma abordagem automática para a extração de correlações litológicas a partir de descrições digitais de poços de exploração. Conforme explicamos no capítulo 5, na tarefa de correlação, o geólogo está interessado em averiguar a continuidade lateral de fácies sedimentares em áreas de subsuperfície. Esta tarefa muitas vezes é realizada manualmente, seja por falta de um sistema que auxilie na sua resolução, seja por falta de uma estruturação formal dos dados de reservatórios coletados por geólogos. A ontologia de domínio desenvolvida por Lorenzatti (2009) e incorporada no software Strataledge®[1], tornou possível que geólogos realizem descrições de fácies sedimentares de maneira uniforme, com um vocabulário formal bem definido, evitando ambiguidades e diferenças resultantes de estilos de escrita distintos. O Strataledge® é um sistema para a descrição de testemunhos baseado em ontologias a partir de uma interface gráfica. Este software permite que as descrições sejam feitas com uma grande riqueza de detalhes, pois sua ontologia foi concebida com o auxílio de diversos especialistas, buscando especificar o conjunto de atributos descritivos de fácies sedimentares que é relevante para os geólogos. Este trabalho é pioneiro em buscar a interpretação automática dos dados gerados a partir desta abordagem.

Neste capítulo, apresentaremos a forma de aquisição de nossos dados, seu padrão, seus atributos e o tratamento a qual foram submetidos. Além disso, apresentaremos conceitualmente nossa abordagem.

8.1 Aquisição de Dados de Teste

Um dos grandes desafios deste trabalho foi encontrar dados que estivessem disponíveis para realizarmos experimentos. Dados geológicos frequentemente possuem importância estratégica para as empresas que os possuem, e consequentemente costumam ser sigilosos. As descrições aqui utilizadas são provenientes de duas fontes: Companhia de Pesquisa de Recursos Mineiras (CPRM)[2] do Rio Grande do Sul; e dados publicados no trabalho de Rodrigues (2010). As descrições obtidas da CPRM (Figura 8.1) correspondem a campanhas de carvão na região de Osório, no estado do Rio Grande do Sul, na formação Rio Bonito, na Bacia do Paraná (Figura 8.3), enquanto as descrições contidas no trabalho de Rodrigues correspondem a poços de exploração de petróleo (Figura 8.2), pertencentes ao campo de Santa Luzia, na bacia do Espírito Santo (Figura 8.4). Ambas descrições foram validadas por geólogos e introduzidas no sistema Strataledge® pelos mesmos.

[1] `http://www.endeeper.com/products/software/strataledge`
[2] http://www.cprm.gov.br/

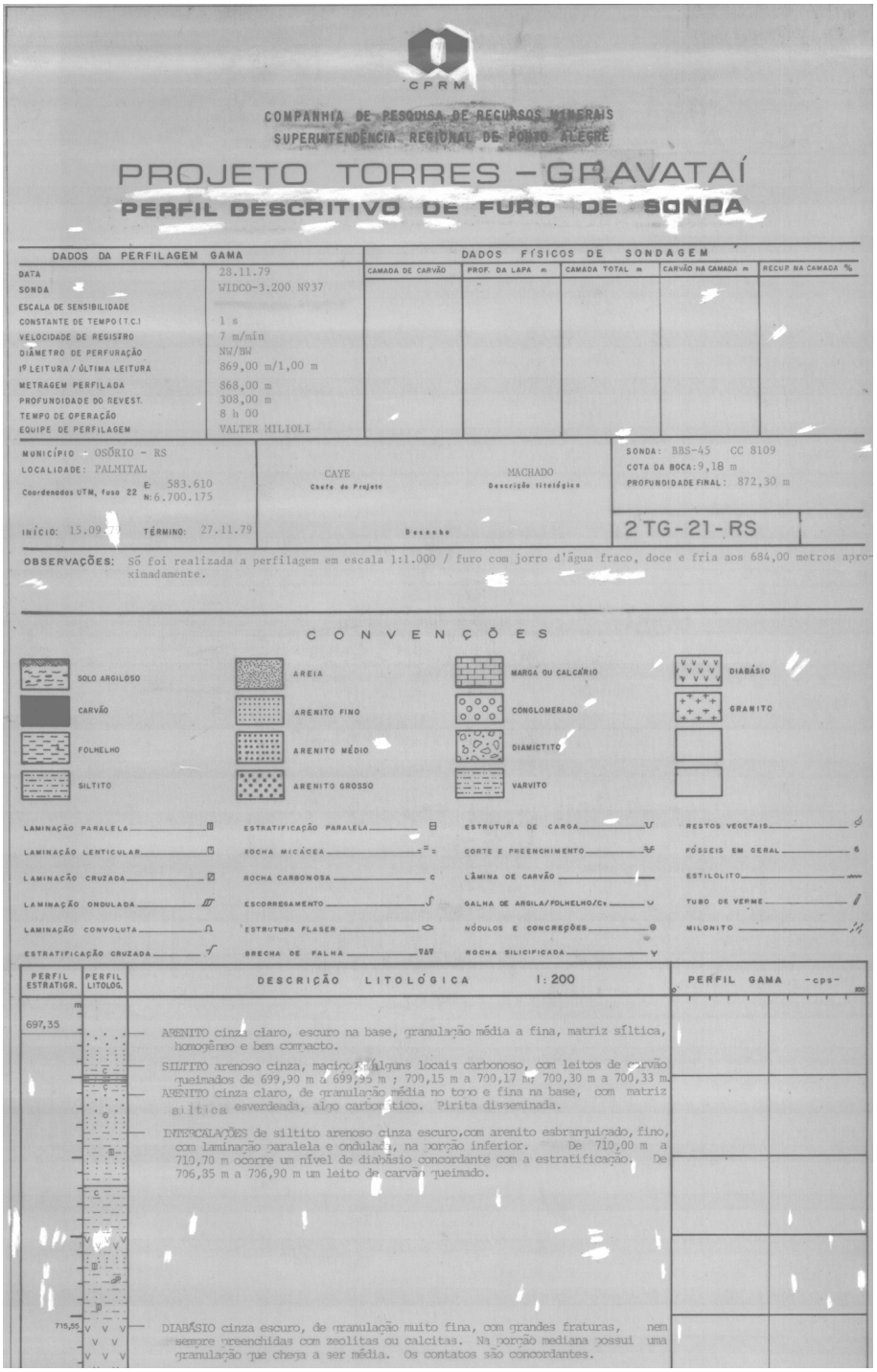

Figura 8.1: Imagem parcial de uma descrição original da CPRM, cujos dados foram utilizados neste trabalho.

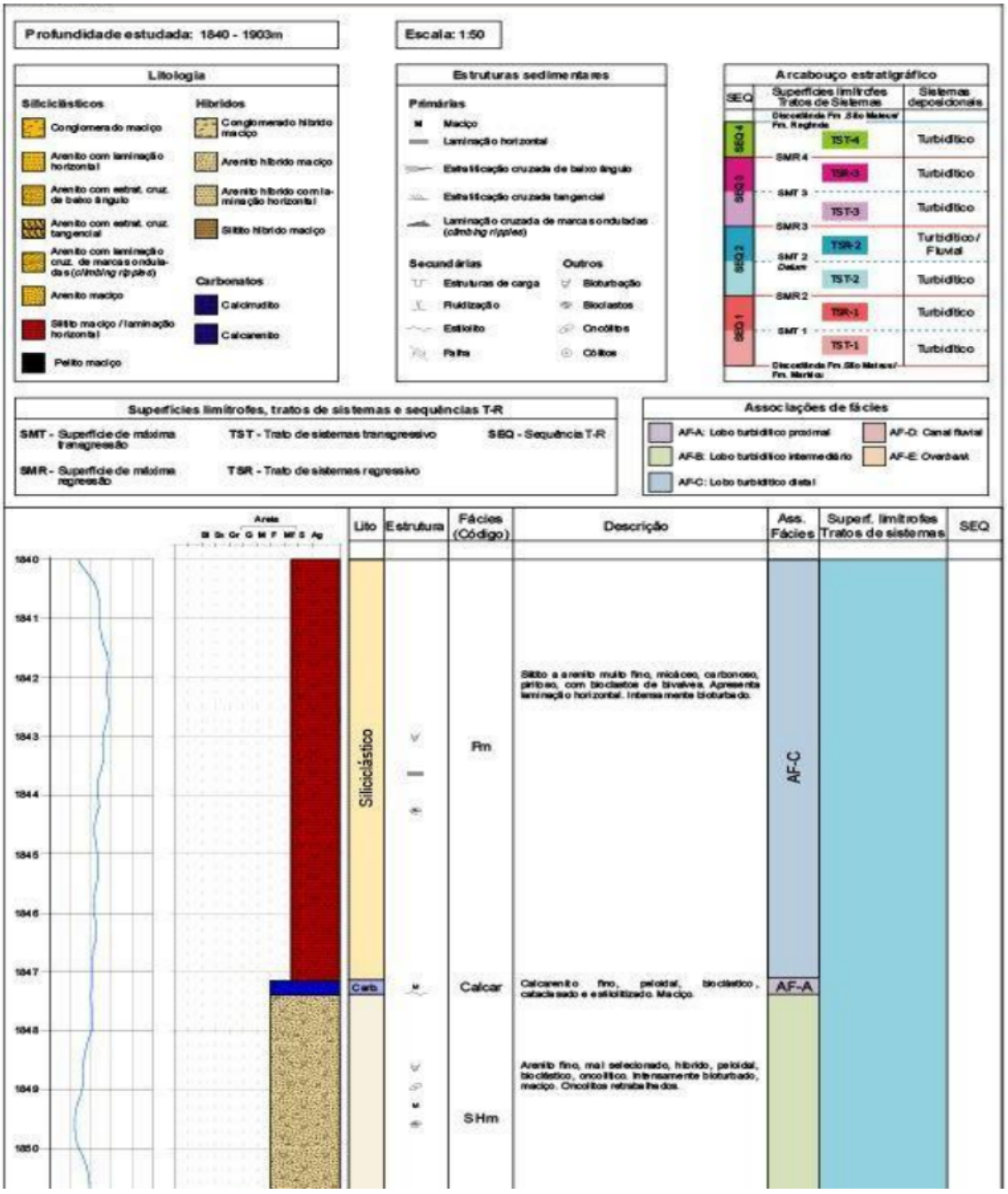

Figura 8.2: Imagem parcial de uma descrição original existente no trabalho de (RODRIGUES, 2010).

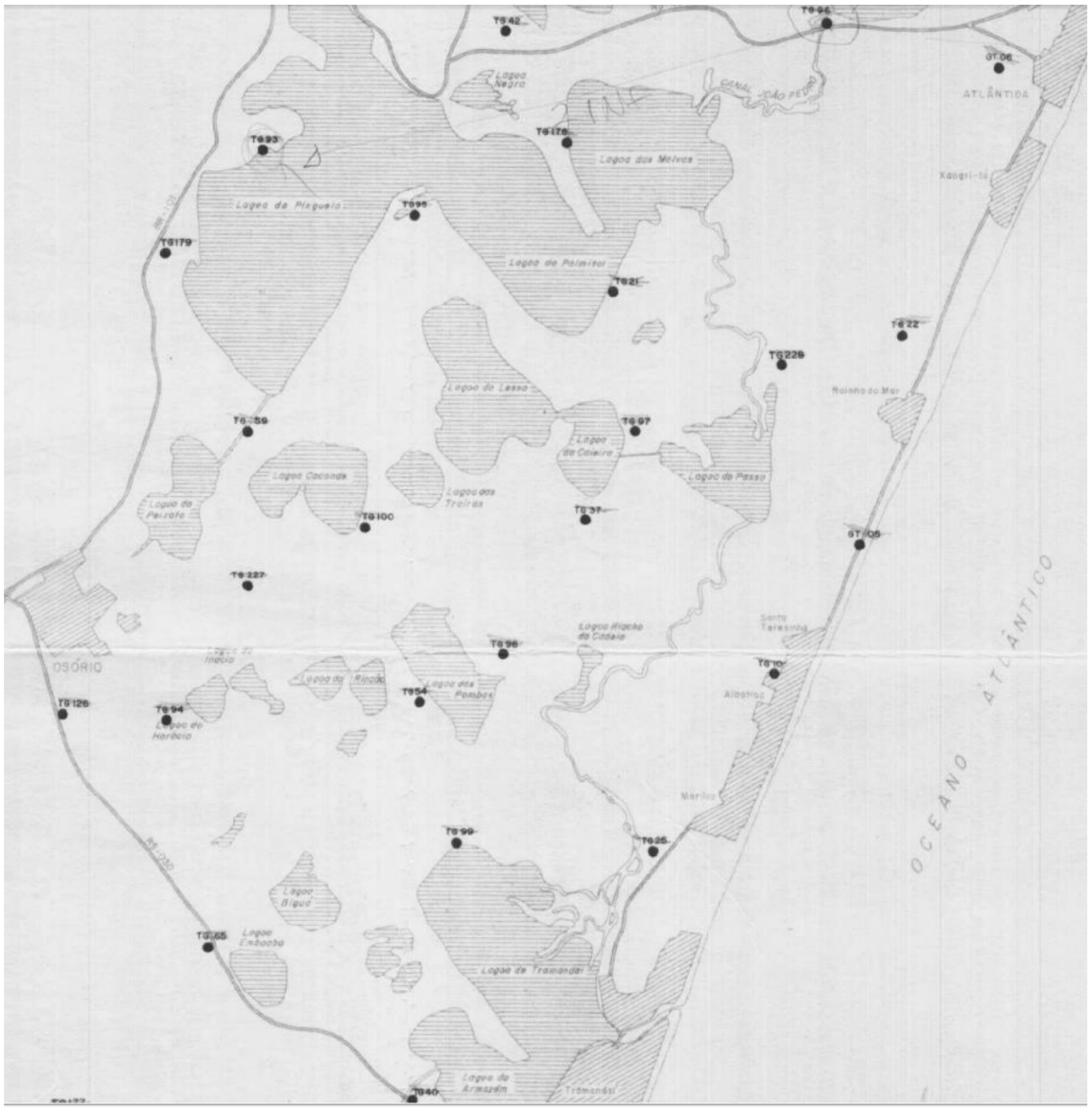

Figura 8.3: Mapa relativo a região de Osório, da formação Rio Bonito. Fornecido pela CPRM.

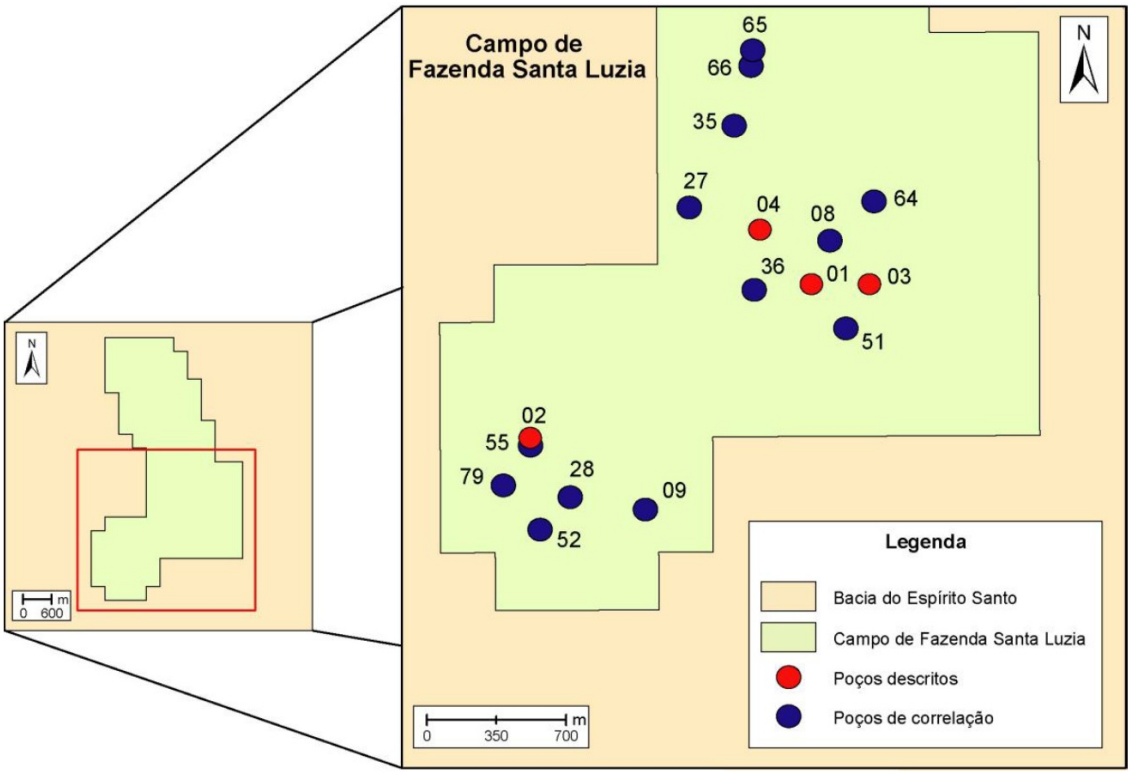

Figura 8.4: Mapa relativo ao campo de Santa Luzia, com os poços descritos marcados em azul. Extraído de (RODRIGUES, 2010).

As descrições realizadas pelos geólogos no Strataledge® são exportadas no formato de um arquivo XML[3] (*Extensible Markup Language*). Nestes arquivos, encontramos informações sobre o poço descrito, como número do poço, localização, etc, e informações sobre as fácies descritas deste poço. No trecho de código 8.1, é possível observar a estrutura em que as informações referentes à fácies sedimentares são armazenadas.

```
<Contact>
      <ContactType/>
</Contact>
<Facies FaciesEmpty="false">
      <FaciesTopMeasure/>
      <FaciesBottomMeasure/>
      <FaciesRockClass/>
      <GrainSize/>
      <Name/>
      <Lithology/>
      <StructureMAIN/>
      <Structure/>
      <Sorting/>
      <Roundness/>
      <Sphericity/>
</Facies>
```

[3]http://www.w3.org/TR/REC-xml/

```
<Contact>
      <ContactType/>
</Contact>
```

Trecho de Código 8.1: Exemplo de arquivo XML.

As tags *«Contact>* representam o tipo de contato existente entre duas fácies, e estão descritos fora das tags *<Facies>* porque compartilham o mesmo valor para ambas. A propriedade *FaciesEmpty* serve para controle, pois podemos ter perdas de material de descrição e, desta maneira, geólogos podem representar esta perda com fácies vazias. Além disto, nas tags *FaciesTopMeasure* e *FaciesBottomMeasure* é descrita a medida de profundidade de uma fácies. As demais tags representam propriedades de fácies cujos conceitos foram apresentados no capítulo 5.

Infelizmente as fácies existentes nestas descrições que utilizamos nem sempre possuem todos seus atributos descritos. As causas são várias: o foco do geólogo era outro que não a correlação, perda de material das amostras de rocha durante a sondagem, falta de meios e ferramentas para o geólogo descrever, etc. Este fato obviamente prejudica a obtenção de resultados.

8.2 Proposta

Uma vez que as descrições de fácies obedecem à estrutura formal estabelecida por uma ontologia de domínio, podemos recuperá-las de acordo com inúmeros atributos e também podemos submetê-las ao processamento computacional. Além disso, dado que a ontologia de domínio especifica uma conceitualização compartilhada pela comunidade a respeito dos objetos do domínio, podemos lidar com dados descritos por geólogos de diferentes escolas, minimizando os problemas da subjetividade e estilo pessoal de descrição.

Neste trabalho, assim como o trabalho de Waterman et al (1987), adotamos a suposição de que há uma equivalência formal entre o problema de correlação litológica, na Geologia, e o problema de alinhamento de sequências de DNA, na Bioinformática. Esta constatação é resultante de diversas entrevistas com especialistas, realizadas com o intuito de compreender a maneira que geólogos realizam a correlação.

A equivalência entre as duas tarefas dá-se quanto ao seu objetivo. No alinhamento de sequências, queremos encontrar subsequências de DNA, RNA ou de proteínas semelhantes entre duas ou mais sequências. Na correlação de poços, queremos encontrar subsequências de fácies sedimentares semelhantes em dois ou mais poços de petróleo.

Deste modo, propomos a utilização de versões modificadas de algoritmos de alinhamento de DNA para realizar a correlação litológica. Esta escolha é motivada pela grande maturidade destes algoritmos, em relação às outras abordagens de correlação litológica encontradas na literatura.

Um desafio que deve ser enfrentado para realizar a adaptação dos algoritmos de alinhamento para a tarefa em foco neste trabalho está no fato de que algoritmos de alinhamento de DNA realizam comparações triviais entre caracteres. Esta característica constitui um desafio porque os objetos geológicos que são alinhados na correlação litológica são objetos complexos, compostos por diversos atributos, muitos deles qualitativos, onde comparações simples não são suficientes.

Para solucionar o problema de comparações de fácies, propomos utilizar técnicas de agrupamento de dados. Isto nos permite realizar comparações com abstrações das fácies,

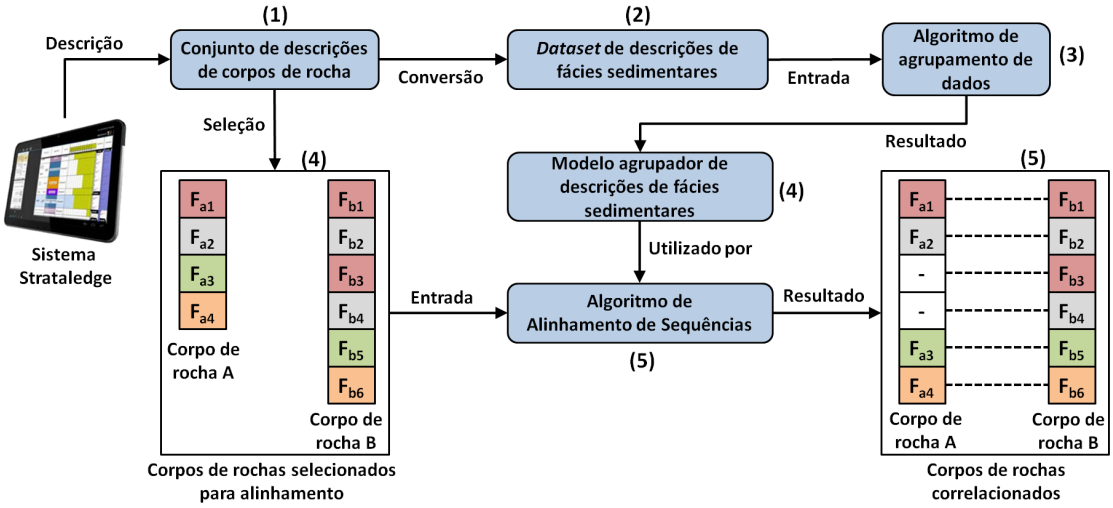

Figura 8.5: Workflow do sistema

que são os *clusters* gerados pelo algoritmo. Sendo assim, duas fácies são consideradas semelhantes caso pertençam a um mesmo *cluster*, e consideradas distintas caso pertençam a *clusters* diferentes. Com isto, reduzimos o problema de comparação de objetos complexos a uma comparação de objetos simples, pois podemos representar *clusters* de maneira numérica. Agora, com comparações simples, podemos realizar alinhamentos nos poços de maneira bastante similar a que se realizam alinhamentos de sequências genéticas.

A nossa abordagem é representada esquematicamente na Figura 8.5 e compreende as seguintes etapas:

1. Geólogos descrevem um conjunto testemunhos de rocha, adquiridos durante a perfuração de poços, utilizando o sistema Strataledge®.

2. Convertemos os dados para um dataset de entrada para um algoritmo de agrupamento.

3. Treinamos o agrupador e geramos um modelo de agrupamento que utilizaremos para comparar as instâncias de fácies sedimentares.

4. Escolhemos os poços que desejamos correlacionar e que, junto com o modelo agrupador são utilizados como entrada para o algoritmo.

5. O resultado é um alinhamento equivalente a uma proposta de correlação litológica para estes poços.

No capítulo seguinte (Capítulo 9), propomos uma implementação para a abordagem apresentada neste capítulo. Justificamos nossas escolhas de algoritmos, explicamos como convertemos nossas descrições, os critérios e parâmetros utilizados para gerar modelos de agrupamento, além da interface e funcionalidades de nossa implementação.

9 UM SISTEMA PARA CORRELAÇÃO LITOLÓGICA

Existem diversas escolhas a serem realizadas para implementarmos a abordagem proposta no capítulo 8. Entre elas, estão os algoritmos de alinhamento e de agrupamento a serem utilizados, além de seus parâmetros e a forma que serão adaptados para o domínio da Geologia. Neste capítulo, apresentaremos nossas escolhas (Seção 9.1) e a filtragem e conversão dos dados apresentados na seção 8.1 para um *dataset* de entrada compatível com o algoritmo de agrupamento (Seção 9.2). A Seção 9.3 apresenta a maneira que geramos modelos de agrupamento e a Seção 9.4 descreve o modo que representamos os dados em nosso sistema, bem como a interface e suas funcionalidades e os parâmetros configuráveis pelo usuário.

É importante salientar que o sistema implementado pode ser compreendido como uma plataforma de pesquisa para abordagens automáticas para alinhamento litológico, permitindo que outros algoritmos de alinhamento (NEEDLEMAN; WUNSCH, 1970) ou outras abordagens de agrupamento sejam integrados e utilizados no futuro.

9.1 Escolhas Iniciais

Primeiramente, foi necessário selecionar um algoritmo de alinhamento de sequências. O algoritmo selecionado foi o de Smith-Waterman (Seção 3.1). São quatro os principais motivos para termos realizado esta escolha:

- O algoritmo de Smith-Waterman encontra alinhamentos locais entre sequências, e no domínio da Geologia também queremos encontrar similaridade entre sequências locais, pois frequentemente os poços não são similares em toda sua extensão.
- É um algoritmo bastante maduro, sendo um dos mais citados neste domínio.
- Os alinhamentos encontrados são ótimos, com relação aos parâmetros de entrada.
- O algoritmo pode encontrar mais de um alinhamento ótimo, possibilitando que o usuário escolha o que melhor lhe favoreça.

Para realizar o agrupamento, o algoritmo que escolhemos foi o EM (Seção 4.1). Esta escolha foi realizada principalmente pelo fato de não termos conhecimento a priori do número de *clusters* a serem encontrados. A abordagem permite utilizar outros algoritmos de agrupamento, porém o EM resultou em agrupamentos considerados satisfatórios por especialistas. Em vez de implementarmos o algoritmo EM, optamos por utilizar uma biblioteca de algoritmos de aprendizado de máquina chamada Weka (HALL et al., 2009). Além da implementação do algoritmo que a ferramenta fornece, também pudemos

aproveitar outros benefícios, como visualizadores de modelos e dados estatísticos dos *datasets*. Além disto, o Weka possui o arquivo ARFF como uma entrada de dados padrão para todos seus algoritmos, com tratamento para instâncias que possuam valores de atributos nulos, fato recorrente em nossos dados.

A seguir, apresentaremos a estrutura de um arquivo ARFF, bem como a conversão dos dados apresentados na seção (8.1), para um formato no padrão deste arquivo.

9.2 O Arquivo ARFF

O arquivo ARFF é uma maneira padrão de representar *datasets* com instâncias independentes, não ordenadas e não relacionadas (WITTEN; FRANK; HALL, 2011). Um arquivo ARFF é composto por atributos e por instâncias. O arquivo inicia com o identificador *@relation*, seguido do nome do *dataset*. Para declarar um atributo inicia-se uma linha com o identificador @ seguido da palavra-chave *attribute*. Os atributos podem ser nominais, numéricos, strings ou datas, porém no contexto deste trabalho consideraremos apenas atributos nominais e atributos numéricos. *Atributos Nominais* devem ter seus possíveis valores declarados entre um par de chaves, logo após a declaração do nome do atributo. *Atributos Numéricos* podem assumir qualquer valor real, e em sua declaração o nome do atributo é seguido da palavra-chave *numeric*. As instâncias são descritas após o identificador *@data*. Cada linha do arquivo representa uma instância, e os valores de seus atributos são separados por vírgulas. A ordem dos valores é a mesma ordem utilizada na descrição dos atributos. Valores de atributos vazios são representados com o símbolo *?*. No trecho de código 9.1, apresentamos um exemplo de arquivo com 7 atributos e 3 instâncias.

```
@relation facies

@attribute grainSize {CoarseGrained, MediumGrained}
@attribute rockColor {White, Red, Black}
@attribute lithology {Coal, Sandstone}
@attribute mainStructure {Bioturbation, Massive}
@attribute sorting {VeryWellSorted, VeryPoorlySorted}
@attribute roundness {VeryAngular, WellRounded}
@attribute sphericity {High, Medium}

@data
CoarseGrained, Black, Coal, Massive, ?, ?, ?
FineGrained, ?, Standstone, ?, ?, VeryAngular, Medium
FineGrained, Black, Coal, Fractured, ?, ?, High
```

Trecho de Código 9.1: Exemplo de arquivo ARFF.

A conversão dos arquivos XML, exportados do software Strataledge®, para um arquivo ARFF é feita de forma automática. Uma divisão prévia dos dados, também automática, é realizada: *Rochas Sedimentares* e *Rochas Ígneas* ou *Metamórficas* são separadas em grupos diferentes. Isto ocorre porque rochas com processos de formação diferentes não podem ser consideradas similares em uma correlação litológica. Como esta separação é conhecimento fundamental de domínio, optou-se por realizar a separação dos dados antes de submetê-los ao algoritmo de agrupamento.

Para realizar a conversão, inicialmente realizamos a leitura do conjunto de poços que formarão o *dataset*, onde cada instância de fácies sedimentar f é convertida para um vetor de atributos v_f, formando um conjunto V de vetores de atributos. Neste vetor v_f, cada posição p representa um atributo a, descrito na ontologia, em que está armazenado um valor referente a instância da fácies f descrita. Para representar estruturas sedimentares foram necessárias algumas modificações. Cada vetor v_f possui uma posição e, que armazena um valor nominal que representa o tipo da sua estrutura deposicional. Porém, fácies sedimentares possuem ainda um número arbitrário de estruturas secundárias, o que impossibilita a criação de um número fixo de posições neste vetor para representá-las. Como solução para este problema, criamos um *attribute* binário para cada tipo de estrutura sedimentar secundária que ocorre no conjunto de fácies, representando a sua ocorrência ou não na fácies f representada por este vetor.

Com este conjunto V de vetores é possível criar um arquivo ARFF de forma simples. Cada posição do vetor é transformado em um *@attribute* e cada vetor é transformado em uma instância do *dataset*. Antes, porém, realizamos alguns tratamentos sobre os dados. As medidas de profundidades das fácies são desconsideradas para o agrupamento, pois a semelhança entre as fácies deve ser avaliada independente de sua profundidade. Além disto, qualquer atributo que não tenha nenhum valor descrito não é inserido no *dataset*, pois não tem representatividade alguma. As estruturas que são tornadas em atributos são apenas as ocorrentes no conjunto V, ou seja, estruturas que não foram descritas neste conjunto não se tornam atributos. Na figura 9.1, representamos um exemplo de conversão de uma fácies em um vetor de características, onde não são consideradas estruturas secundárias.

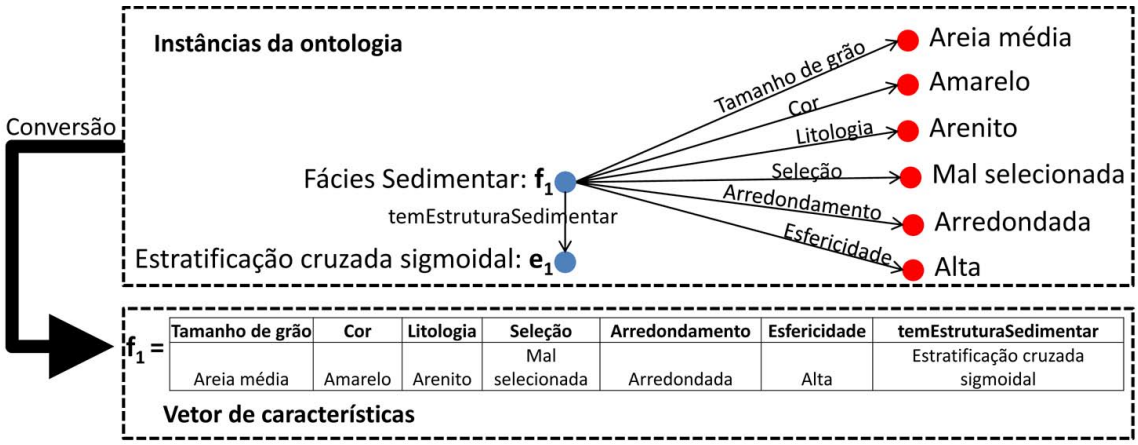

Figura 9.1: Conversão de uma instância de fácies sedimentar em um vetor de características. Figura meramente ilustrativa.

9.3 Modelos de Agrupamento

Com os *datasets* em formato ARFF, foi possível gerar os modelos de agrupamento que são utilizados como entrada do algoritmo de alinhamento de sequências adaptado. Para gerar os modelos de agrupamento do algoritmo EM utilizamos as funcionalidades da *API* do Weka. Os parâmetros de entrada do algoritmo são o número total de iterações, o desvio padrão mínimo permitido, o número de *clusters* desejado e uma semente para o gerador de

número randômicos. Não existem valores para estes parâmetros que funcionem bem para todos os *datasets*, portanto foi necessário realizar diversos testes até encontrar valores aceitáveis.

Considerando o nosso *dataset*, para o número total de iterações constatamos que com um valor de 1000 iterações já não se tem ganhos representativos nos resultados. Como desvio padrão mínimo escolhemos o valor 0,000001, pois desta maneira o algoritmo não fica limitado pelo valor de desvio. Não definimos um valor total de *clusters* a serem encontrados, de forma que o próprio algoritmo deveria encontrar tal número. Infelizmente não há maneira simples de se escolher uma semente para um gerador randômico, e para este algoritmo, valores iniciais distintos resultarão em modelos distintos. Por esta razão escolhemos uma abordagem iterativa. Geramos modelos para sementes que variam do número 1 até o número 2000, e armazenamos os modelos que satisfazem critérios escolhidos previamente. Os principais critérios utilizados são o maior e o menor valor de verossimilhança e o maior e o menor número de *clusters* dos modelos. Foram gerados modelos tanto para as instâncias de *Rochas Sedimentares*, quanto para as instâncias de *Rochas Ígneas* ou *Metamórficas*.

De acordo com Guyon (2009), algoritmos de agrupamentos não podem ser avaliados de forma independente de domínio. Para definir um agrupamento como bom ou ruim devemos levar em consideração o objetivo final. Por este motivo, a qualidade dos modelos aqui gerados pode ser avaliada de duas maneiras. A primeira maneira é avaliando se os *clusters* encontrados no modelo possuem alguma relevância geológica. Neste caso, é possível excluir modelos que não reflitam características da realidade. A segunda maneira é avaliar o resultado final do sistema, ao invés de avaliar os *clusters*. Desta forma, deveríamos avaliar a qualidade das correlações geradas pelo sistema. Ambas avaliações requerem conhecimento sobre o domínio, e portanto, necessitam de especialistas para poderem ser realizadas.

9.4 Implementação do Sistema

Nesta seção, apresentaremos primeiramente como representamos os poços em nosso sistema. Em seguida, as funcionalidades existentes, a interface, os parâmetros possíveis e exemplos de correlações litológicas obtidas com nossa implementação.

Os poços descritos pelos geólogos são representados em nossa implementação pela classe *StratigraphicProfile*. Cada objeto de poço possui um atributo do tipo string *profileName*, para armazenar seu nome, e uma lista de objetos do tipo *SedimentaryFacies*. A classe *SedimentaryFacies* representa as instâncias de fácies sedimentares descritas pelos geólogos para aquele poço, e possui 14 atributos, listados a seguir.

topMeasure: Atributo numérico que representa a profundidade inicial em que fácies sedimentar é descrita.

bottomMeasure: Atributo numérico que representa a profundidade em que a descrição da fácies termina.

lithology: Possui o tipo *Lithology*, e representa a composição litológica de uma fácies.

faciesName: É um atributo do tipo string, que armazena o campo equivalente no arquivo de descrição, não é utilizado para realizar a correlação.

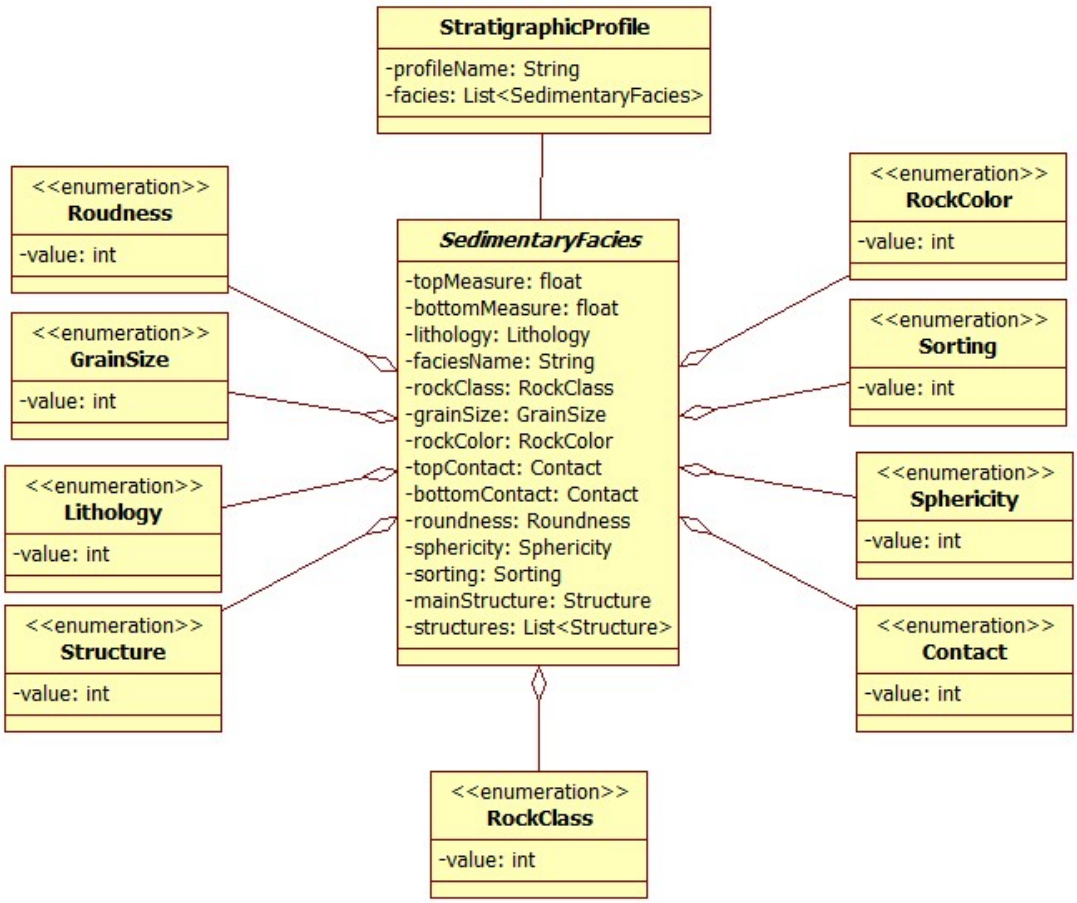

Figura 9.2: Representação das descrições de poços em um diagrama de classes.

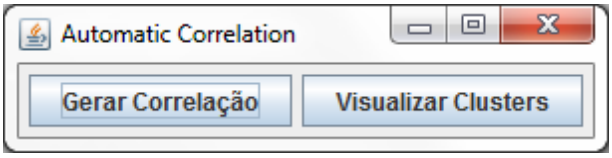

Figura 9.3: Tela inicial do sistema.

rockClass: Representa a classe genética da rocha, que pode ser Sedimentar, Ígnea ou Metamórfica.

grainSize: Atributo para armazenar a granulometria da fácies. É do tipo *GrainSize*.

rockColor: Possui o tipo *RockColor*, representa a cor da instância da fácies sedimentar.

roundness: Representa a propriedade de arredondamento de uma fácies sedimentar. É do tipo *Roundness*.

sphericity: Atributo de tipo *Sphericity*, que representa a propriedade de esfericidade de uma fácies.

sorting: Propriedade de seleção de uma fácies. É do tipo *Sorting*.

mainStructure: Representa a estrutura sedimentar dominante da fácies. Este atributo é do tipo *Structure*.

structures: É uma lista de objetos do tipo *Structure*. Representa todas outras estruturas sedimentares que não são a estrutura dominante.

Na figura 9.3, podemos observar duas funcionalidades existentes para o usuário: *Visualizar Clusters*; e *Gerar Correlação*. A funcionalidade de visualização consiste em apresentar para o usuários as instâncias de um *dataset* agrupadas nos respectivos *clusters*, determinados por um modelo específico previamente treinado. A visualização permite que o geólogo avalie um modelo de agrupamento de forma visual. A funcionalidade de correlação consiste em realizar um alinhamento de dois poços descritos de tal forma que exista uma equivalência a uma correlação litológica. Nas subseções 9.4.1 e 9.4.2, apresentamos tais funcionalidades em detalhe.

As funcionalidades de converter *datasets*, a partir de arquivos *XML*, e de gerar *modelos de agrupamento*, a partir de arquivos *ARFF*, ainda não possuem uma interface gráfica implementada.

9.4.1 Visualização de Clusters

Para visualizar os *clusters* gerados por um modelo, primeiramente é necessário que o usuário escolha o *dataset* desejado, no formato *ARFF*, e o próprio modelo de agrupamento, previamente treinado, gerado com a biblioteca *Weka* (Figura 9.4). Um exemplo da visualização de dois *clusters* com suas respectivas instâncias pode ser observado na figura 9.5. As primeiras 11 colunas são utilizadas para representar a granulometria das fácies, iniciando a partir da menor granulometria (Silte) até a maior (Matacão). Nas demais colunas, utilizamos ícones para os atributos Litologia e Estrutura Sedimentar, definidos a partir dos metaconstrutos apresentados no capítulo 7, como forma de representação para o conhecimento visual. Como é possível uma fácies possuir mais

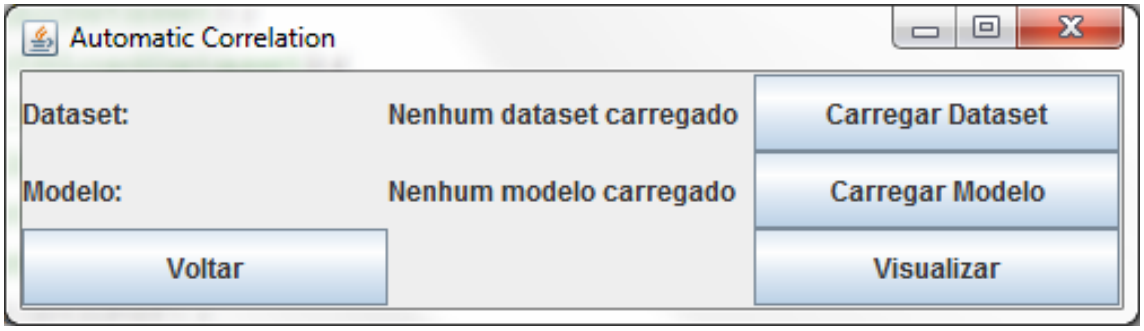

Figura 9.4: Tela de escolha de modelo agrupador e *dataset* para visualização.

de uma Estrutura Sedimentar, o primeiro ícone equivale à sua estrutura deposicional, enquanto os demais ícones, se houverem, equivalem à suas estruturas secundárias. Não há nenhuma ordenação quanto a sequência vertical em que as instâncias de fácies são organizadas, bem como também não existe uma relação entre o tamanho da linha da tabela com o tamanho da fácies descrita.

9.4.2 Gerar Correlação

As correlações litológicas são geradas com esta funcionalidade. Para isto é necessário que o usuário carregue os dois poços desejados, no formato de exportação do Strataledge®, e também os modelos e datasets desejados, tanto de *Rochas Sedimentares* como de *Rochas Ígneas ou Metamórficas*, no modelo da biblioteca Weka. Além disto, é necessário definir os parâmetros de *gap*, *match* e *mismatch*, introduzidos no Capítulo 3, referentes ao algoritmo de alinhamento de *Smith-Waterman*. A interface descrita pode ser visualizada na Figura 9.6.

Os *datasets* se fazem necessários, utilizamos a biblioteca Weka para realizar as comparações de fácies sedimentares. Os métodos disponibilizados por esta biblioteca exigem que sejam fornecidos não apenas o *modelo de agrupamento*, mas também os *dataset* que gerou tal modelo.

Na figura 9.7, o leitor pode visualizar um exemplo de correlação obtida através desta implementação da abordagem, realizada com dados da Formação Santa Luzia. As descrições dos poços estão situadas nos cantos esquerdo e direito da janela. As linhas azuis (nos dois tons) representam instâncias que foram alinhadas. Neste exemplo, apenas os atributos de granulometria, litologia e estrutura sedimentar estão visíveis.

Todos alinhamentos ótimos são apresentados ao usuário, ou seja, caso exista mais de um alinhamento considerado ótimo pelo algoritmo, novas janelas serão apresentadas, cada uma com uma proposta de alinhamento diferente. Desta forma, o usuário pode escolher o resultado que mais lhe agrada.

No próximo capítulo, apresentaremos alguns exemplos de correlações gerados a partir de nossa abordagem.

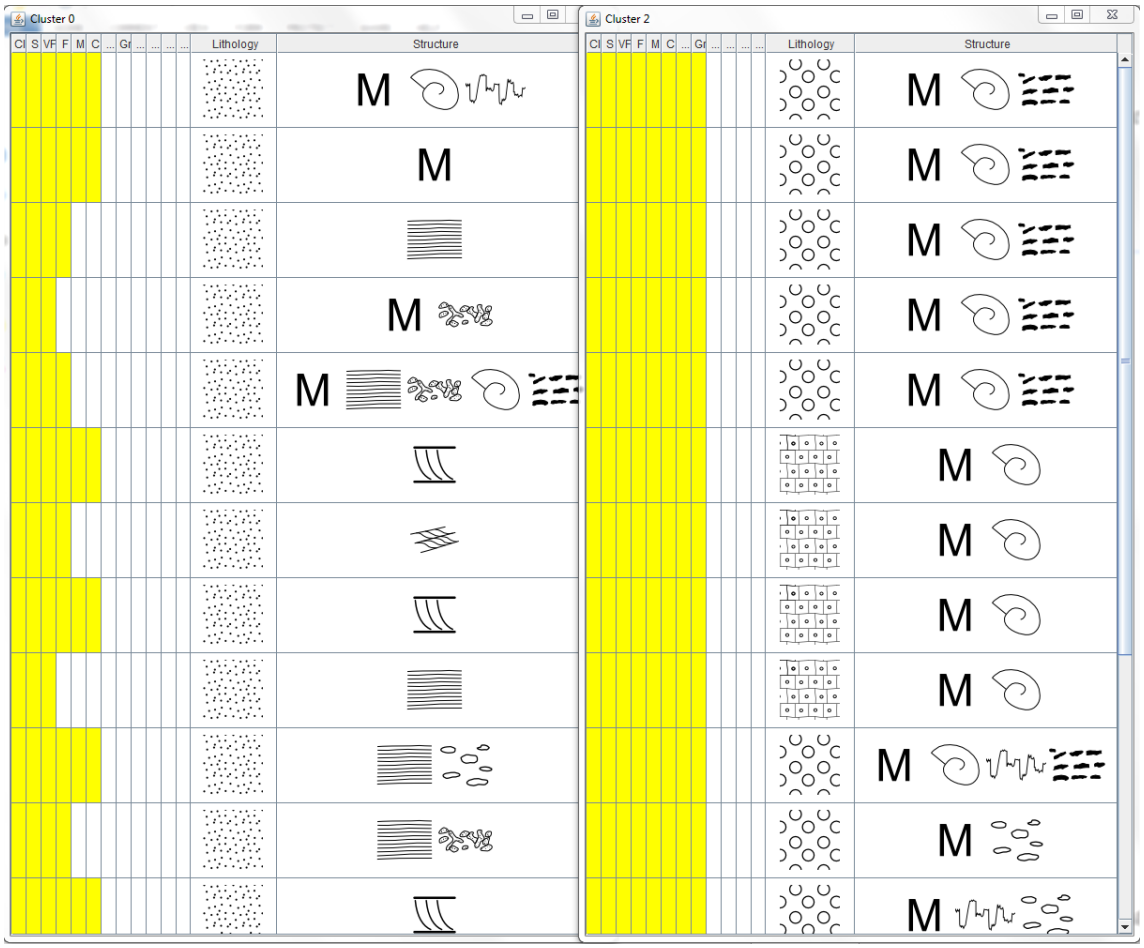

Figura 9.5: Visualização de dois *clusters* existentes em um modelo.

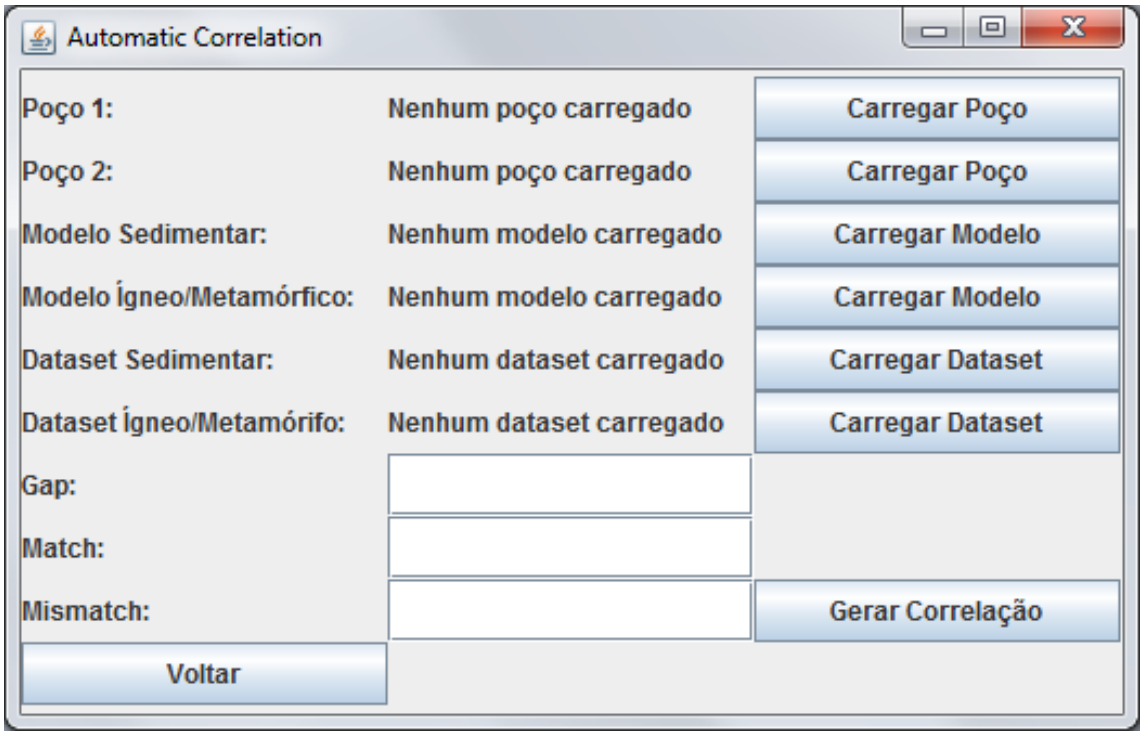

Figura 9.6: Tela para escolha dos arquivos de entradas e parâmetros do algoritmo para gerar uma correlação.

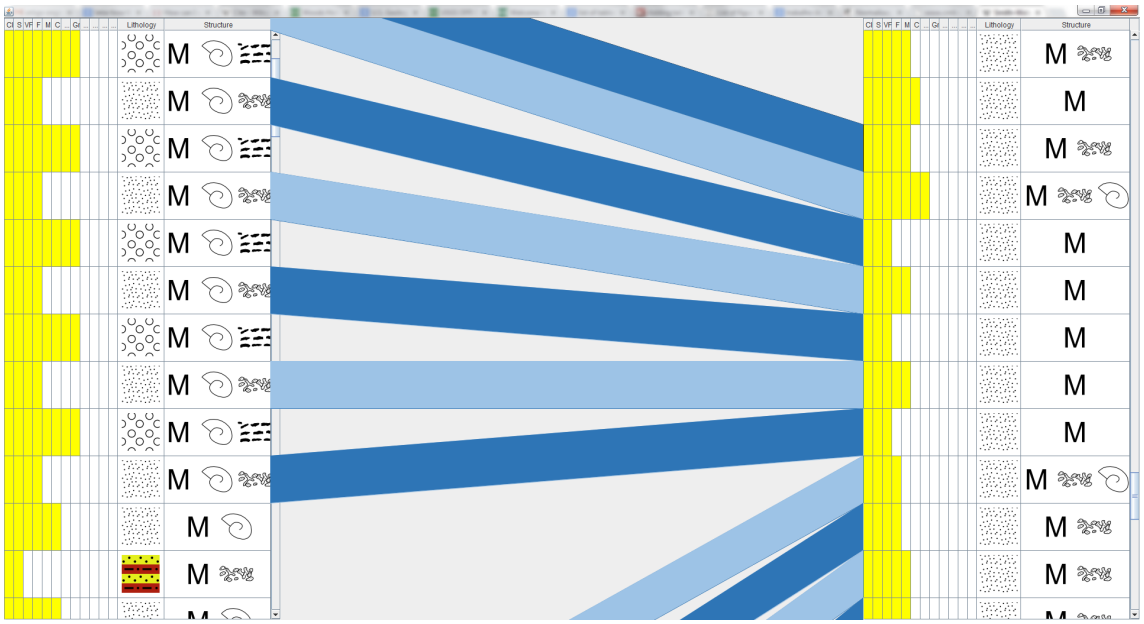

Figura 9.7: Exemplo de correlação gerada por nosso sistema.

10 RESULTADOS

Apresentaremos neste capítulo trechos de exemplos de correlações geradas utilizando nossa plataforma. É importante salientar que os testes apresentados aqui possuem caráter exploratório. A validação da abordagem exige um volume mais representativo de dados reais e o acompanhamento de resultados por um geólogo experiente, etapa que esta prevista em trabalhos futuros.

Os poços utilizados pertencem à formação de Santa Luzia (Figura 8.4), e são os poços de número 1, que possui 90 fácies descritas, e de número 3, que possui 149 fácies descritas. A escolha destes dois poços se deve ao fato de que estão localizados geograficamente próximos, portanto tem maior possibilidade de cortar as mesmas unidades de rocha que podem ser correlacionadas, seus testemunhos possuem baixa perda de material e são os poços com a maior quantidade de atributos descritos. Os poços de números 2 e 4 não serão considerados, pois estão distantes geograficamente e possuem muita perda de material. O exame mais detalhado das descrições mostrou que todos os poços fornecidos pela CPRM foram descritos de modo muito superficial, com poucos atributos qualificados. A falta de detalhamento da descrição limita demasiadamente a extração de correlações, sejam elas automáticas ou feitas manualmente por um geólogo. Por esta razão, os testes com esses poços foram descontinuados e as correlações não serão mostradas neste trabalho.

Nos exemplos, os poços serão apresentados em tabelas, seguindo um modelo semelhante ao da Seção 9.4. O poço 1 está situado à esquerda da janela, enquanto o poço 3 encontra-se no lado direito. As 11 primeiras colunas representam a granulometria das fácies, variando de silte (mais fina) até matacão (mais grossa). A coluna 12 apresenta um ID para cada fácies. Este ID não reflete nenhuma característica geológica e tem intuito apenas de auxiliar o leitor a reconhecer os mesmos trechos entre exemplos diferentes. A coluna 13 apresenta ícones das litologias de cada fácies e a coluna 14 ícones da estrutura deposicional e, quando existente, das estruturas sedimentares secundárias.

Para os testes, variamos os valores dos pesos do algoritmo de Smith-Waterman de *gap*, *match* e *mismatch*. Adotamos a heurística proposta em (SMITH; WATERMAN, 1981), onde um valor de *gap* deve ser pelo menos igual à diferença entre os valores de *match* e *mismatch*. Sendo assim, podemos ter um valor de *gap* que seja igual ou maior do que esta diferença, porém nunca um valor menor. Além disso, também exploramos diferentes modelos de agrupamento, com o intuito de demonstrar os efeitos desta variação nas correlações resultantes.

Escolhemos dois modelos de agrupamento para realizar nossos testes: modelo com maior valor de verossimilhança, e modelo com maior número de *clusters*. Ambos modelos foram gerados através da abordagem descrita na Seção 9.3, onde variamos o valor da

semente para o gerador de números randômicos e armazenamos os modelos que possuíam as melhores medidas avaliadas.

As legendas dos ícones utilizados em nossa plataforma podem ser visualizadas nas Tabelas 10.1 (para litologias) e 10.2 (para estruturas sedimentares).

Litologias		
Arenito 	Conglomerado 	Grainstone 
Heterolito 	Rudstone 	Siltito 

Tabela 10.1: Lista de litologias e seus respectivos ícones.

Realizamos testes com dois conjuntos de parâmetros de *gap*, *match* e *mismatch* para cada modelo de agrupamento. Os parâmetros foram os seguintes:

Gap = 6, *Match* = 3 e *Mismatch* = -3. Estes parâmetros foram escolhidos visando um balanceamento entre os valores de *match* e *mismatch*.

Gap = 13, *Match* = 3 e *Mismatch* = -10. Com estes parâmetros forçamos que ocorram apenas *matchs* de elementos similares.

10.1 Testes com modelo de agrupamento com maior verossimilhança

As Figuras 10.1, 10.2 e 10.3 correspondem a trechos do alinhamento obtido com os parâmetros de *gap* = 6, *match* = 3 e *mismatch* = -3. Já as Figuras 10.4 e 10.5 correspondem a diferentes trechos da correlação obtida com os parâmetros de *gap* = 13, *match* = 3 e *mismatch* = -10.

O primeiro alinhamento (Figuras 10.1, 10.2 e 10.3) correlacionou em sua maioria fácies com litologias de arenito maciço com outras fácies de litologia também de arenito maciço onde se alternam camadas com e sem grânulos. Este resultado possui relevância geológica, visto que tais litologias são correlacionáveis. Neste mesmo alinhamento, também conseguimos identificar uma alternância entre litologias de arenito e litologias de siltito (Figura 10.3), o que também reflete um aspecto geológico coerente. A mudança litológica em que deixam de ocorrer arenitos nos dois poços e começam a predominar siltitos e carbonatos, sem correspondência da litologia nos dois poços determinou o encerramento da correlação, como mostrado na parte de baixo da Figura 10.3.

O segundo alinhamento (Figuras 10.4 e 10.5) não gerou uma correlação com significância geológica. Foram correlacionadas fácies de litologia arenito com fácies de litologia grainstone e rudstone. Este não é um alinhamento correto, visto que arenitos possuem origem siliciclástica (clásticos mobilizados e depositados) , enquanto grainstones e rudstones são carbonáticos (rochas resultantes de precipitação química) e rochas com estas diferentes formações não são correlacionáveis. De acordo com o geólogo, o agrupamento de arenitos e carbonatos neste experimento se deu pela semelhança das duas rochas nesta ocorrência em especial da Formação Santa Luzia. Como os carbonatos da borda da plataforma foram remobilizados e depositados dentro da bacia, adquiriram estruturas de deposição típica de rochas siliciclásticas. Como a diferenciação geologicamente significativa entre as rochas se resumiu a litologia, sendo os demais atributos muito parecidos, não foi possível que o algoritmo de agrupamento

Estruturas Sedimentares		
Bioclasto 	Bioturbação 	Cimento 
Laminação Cruzada 	Estratificação Cruzada 	Falha 
Fluidização 	Fraturado 	Intraclastos 
Estrutura de Carga	Estratificação Cruzada de Baixo Ângulo	Maciço M
Gradação Normal 	Laminação Plano-Paralela 	Gradação Inversa 
Grânulos Dispersos 	Estilolito 	Estratificação Cruzada Tangencial 

Tabela 10.2: Lista de estruturas sedimentares e seus respectivos ícones.

separasse as classes de rochas como um geólogo faria. Para obter este resultado, seria necessário que fossem atribuídos diferentes pesos aos atributos descritos, o que este algoritmo de agrupamento não considera.

Na Figura 10.6, é possível observar um trecho da visualização de um *cluster* gerado no modelo de maior verossimilhança, onde no mesmo *cluster* estão presentes rochas carbonáticas (*grainstones*) e rochas siliciclásticas (arenitos).

10.2 Testes com modelo de agrupamento de maior número de clusters

As Figuras 10.7, 10.8 e 10.9 correspondem trechos do alinhamento obtido com os os parâmetros de *gap* = 6, *match* = 3 e *mismatch* = -3. Já as Figuras 10.10, 10.11 e 10.11, correspondem a diferentes trechos da correlação obtida com os parâmetros de *gap* = 13, *match* = 3 e *mismatch* = -10.

O alinhamento apresentado nas Figuras 10.7, 10.8 e 10.9, também possui problemas ao alinhar litologias geneticamente diferentes (arenitos e ruditos, Figura 10.7), porém mantem a coerência ao correlacionar rochas com granulometrias próximas (tamanhos silte e areia fina), mesma estrutura de acamadamento e presença de estruturas secundárias como grânulos ou gretas de contração. O maior problema deste alinhamento acontece no alinhamento de litologias com diferença granulométrica muito grande (Figura 10.8), apesar de identificar corretamente alternâncias entre as litologias de arenito e siltito (Figura 10.9).

O último exemplo de alinhamento (Figuras 10.10, 10.11 e 10.11) apresenta uma correlação geologicamente satisfatória. Apesar de apresentar algumas inconsistências, como correlacionar granulometrias muito díspares (Figura 10.10), em grande parte do alinhamento foram correlacionadas fácies semelhantes (Figura 10.11), além de ter

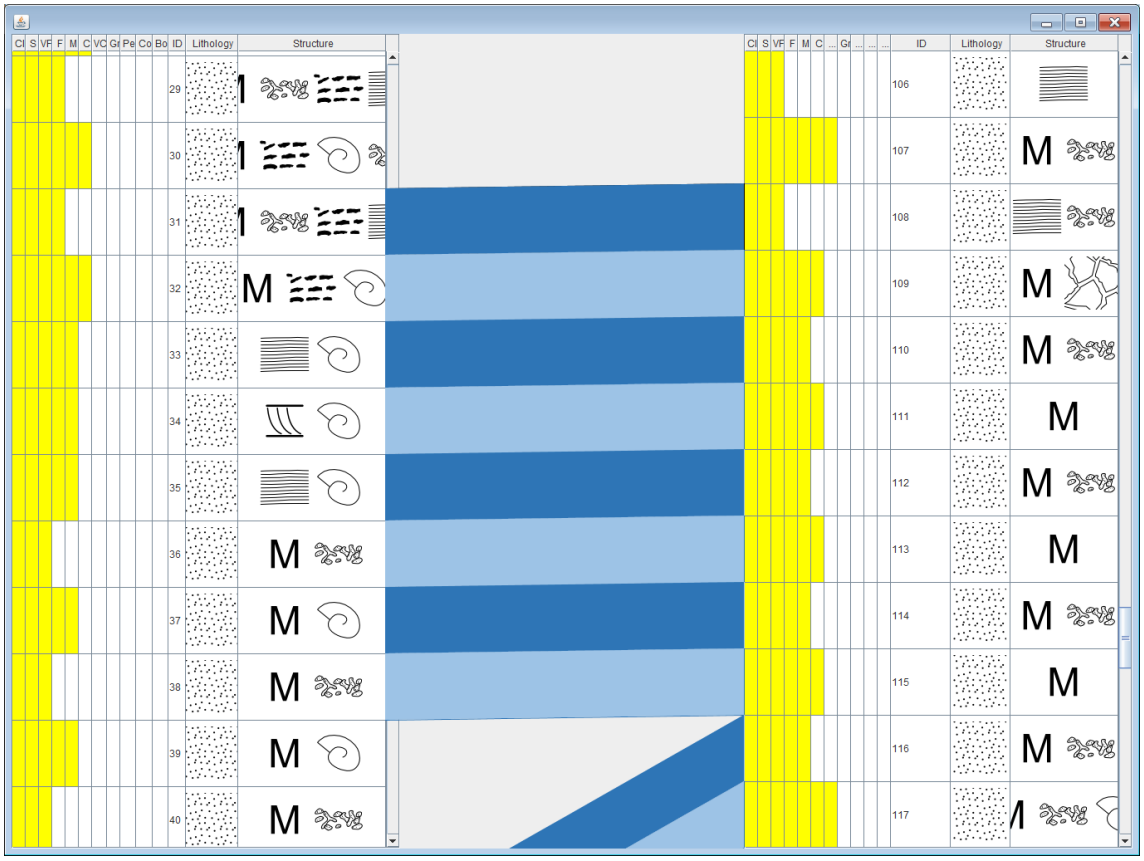

Figura 10.1: Trecho inicial de alinhamento gerado com o modelo de maior valor de verossimilhança e os parâmetros: *gap* = *6*; *match* = 3; *mismatch* = -3.

identificado a alternância de litologias arenito e siltito, com estruturas maciça ou plano-paralela (Figura 10.12). Novamente, o algoritmo acerta ao encerrar a correlação quando a sequência de rochas dos dois poços torna-se diferentes.

Os resultados mostrados nas Seções 10.1 e 10.2, nos levam a acreditar que técnicas de agrupamento de dados não são uma abordagem apropriada para lidar com o problema de comparação de fácies sedimentares. Técnicas de agrupamento de dados parecem ter limitações de aplicação aqui porque geram um julgamento binário (as fácies são ou não similares). No entanto, entrevistas com os geólogos sugerem que o julgamento de similaridade é uma questão de grau, de modo que ele deve variar conforme o grau de similaridade entre as fácies. Além disto, um grande desafio para se trabalhar no domínio geológico é o fato de não existirem *datasets* com grandes amostragens que não possuam uma grande aleatoriedade entre suas instâncias.

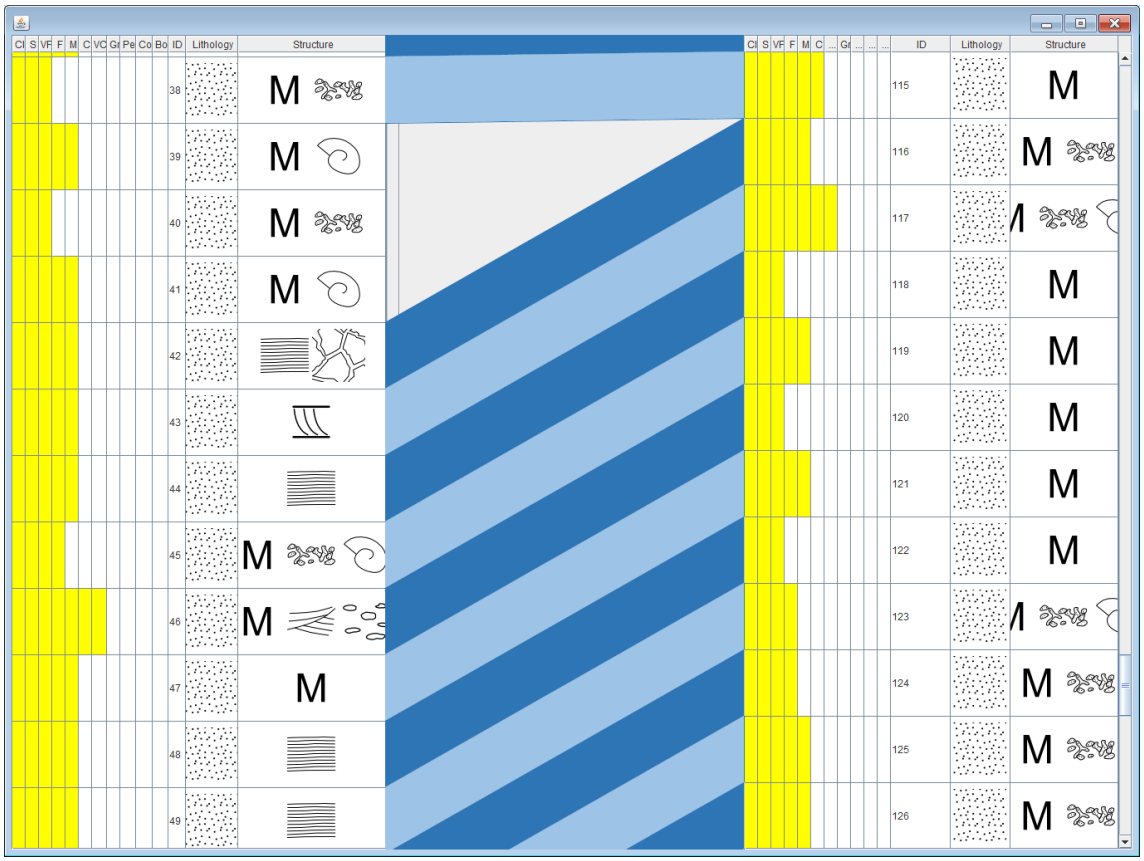

Figura 10.2: Trecho de alinhamento gerado com o modelo de maior valor de verossimilhança e os parâmetros: *gap* = *6*; *match* = 3; *mismatch* = -3.

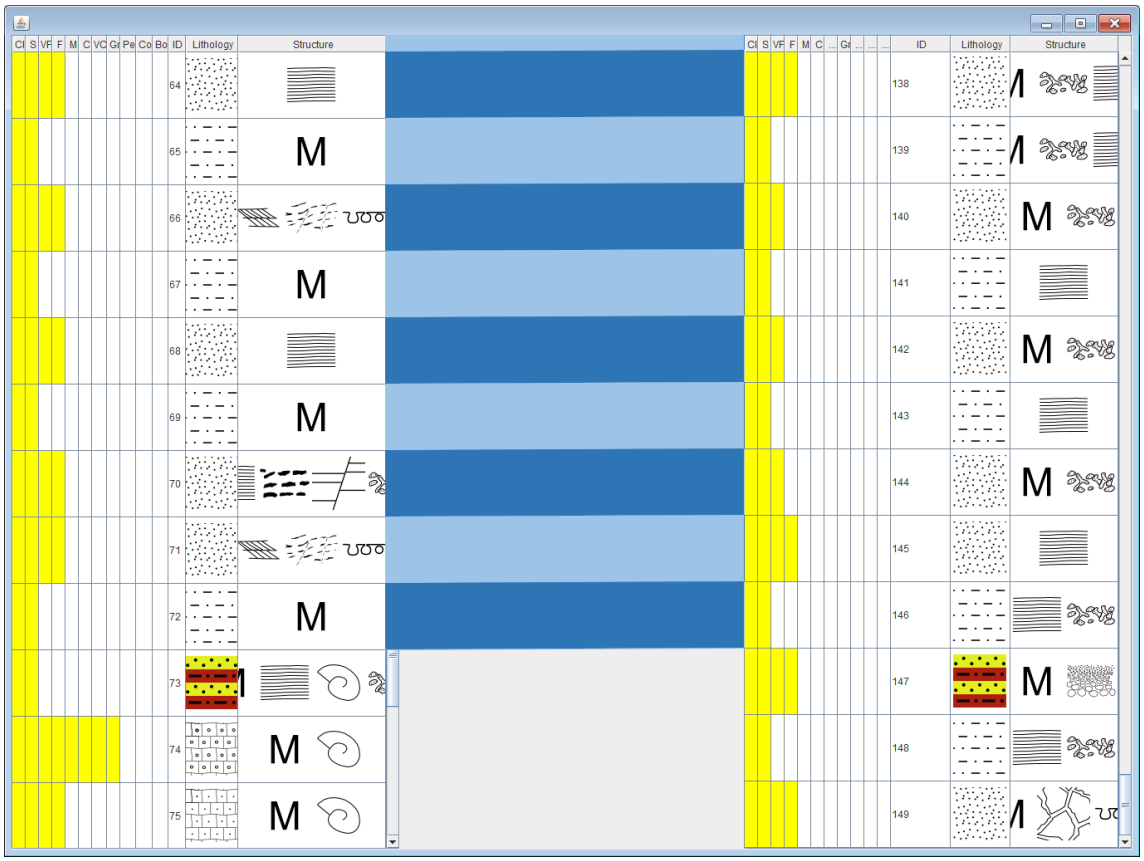

Figura 10.3: Trecho final de alinhamento gerado com o modelo de maior valor de verossimilhança e os parâmetros: *gap* = *6*; *match* = 3; *mismatch* = -3.

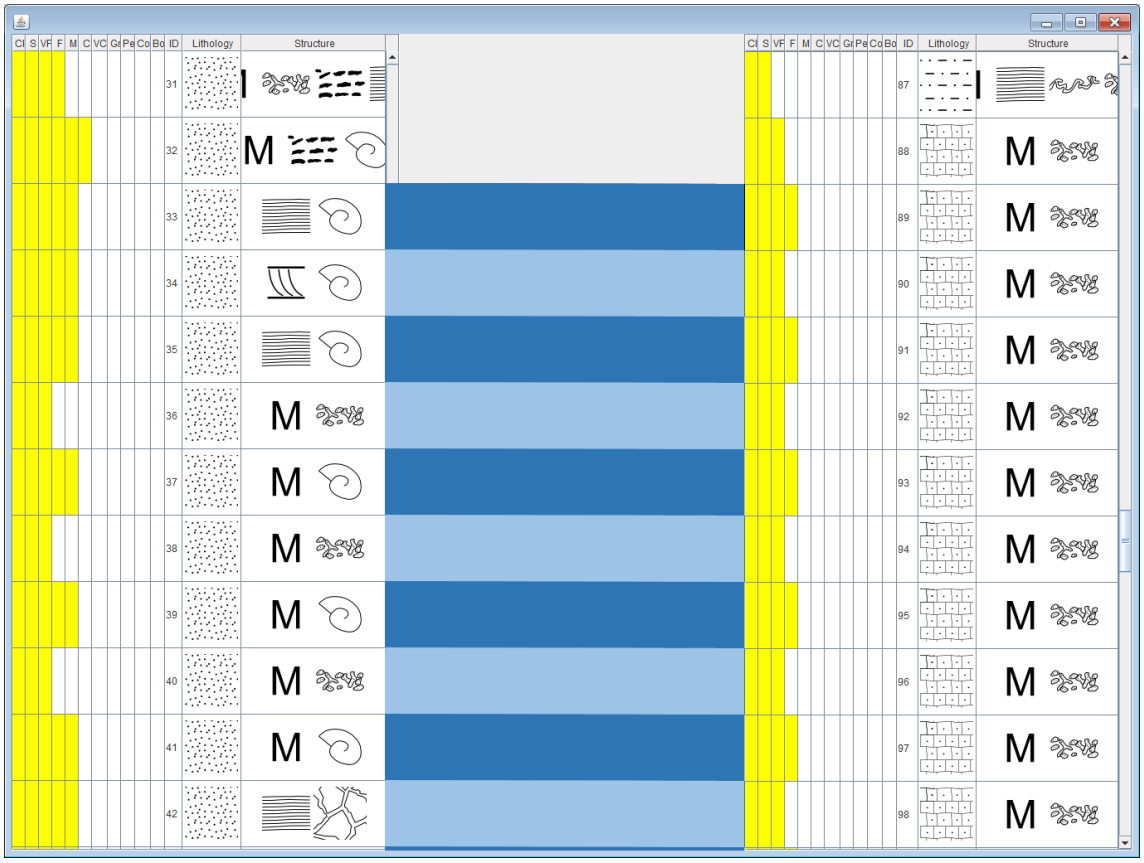

Figura 10.4: Trecho inicial de alinhamento gerado com o modelo de maior valor de verossimilhança e os parâmetros: *gap* = *13*; *match* = 3; *mismatch* = -10.

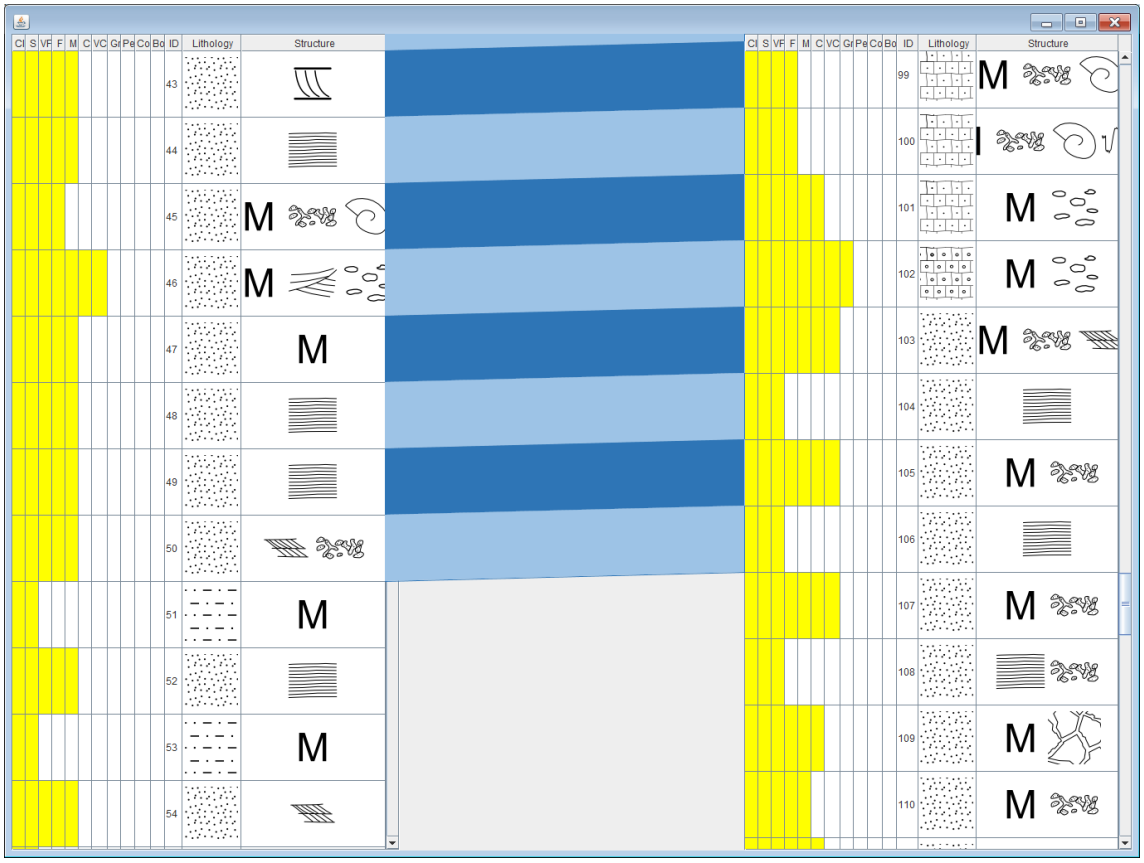

Figura 10.5: Trecho de alinhamento gerado com o modelo de maior valor de verossimilhança e os parâmetros: *gap* = *13*; *match* = 3; *mismatch* = -10.

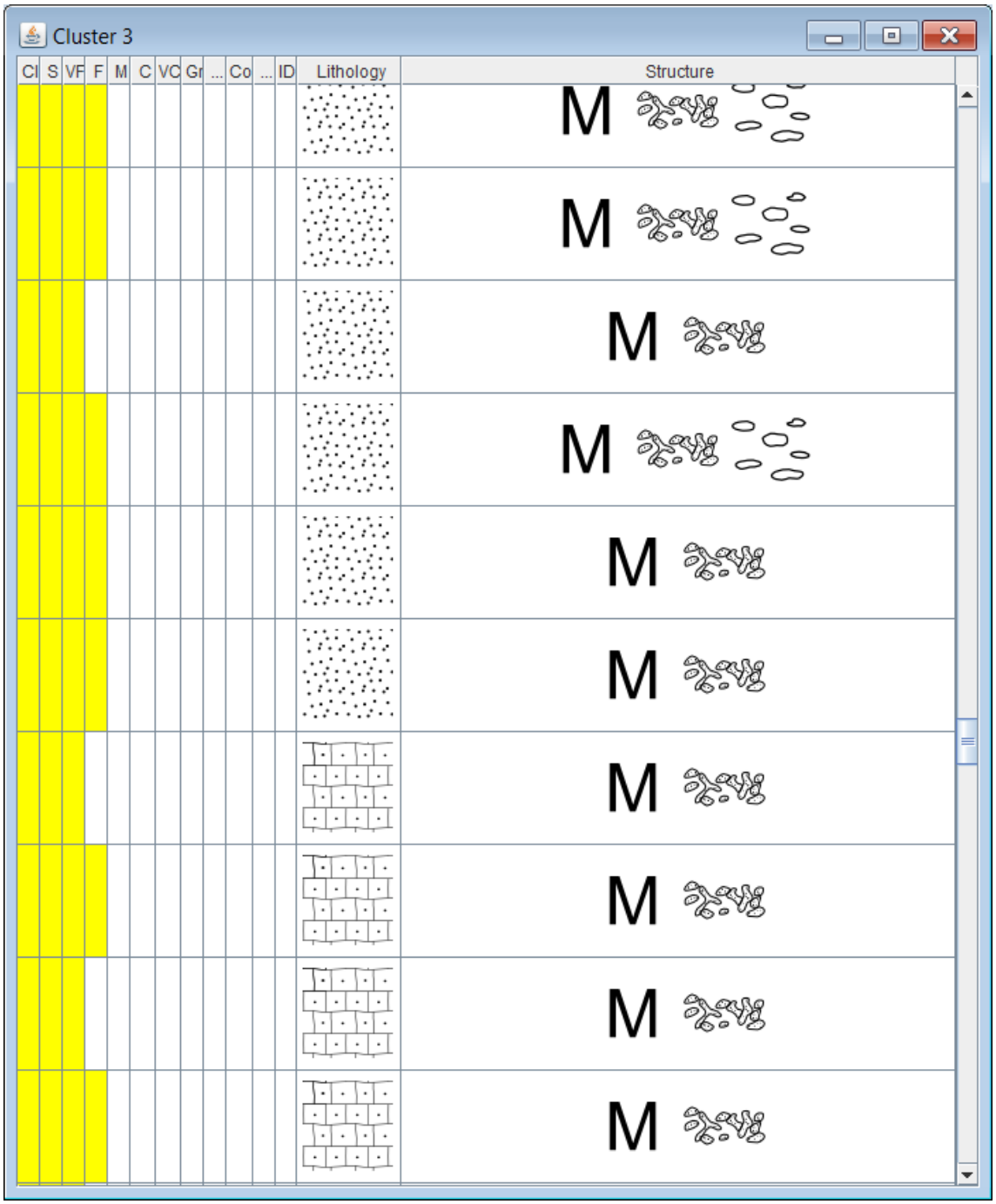

Figura 10.6: Trecho de um *cluster* gerado com o modelo de maior verossimilhança. Nesta imagem, é possível observar uma mistura de fácies sedimentares com formações diferentes e que não deveriam ser agrupadas.

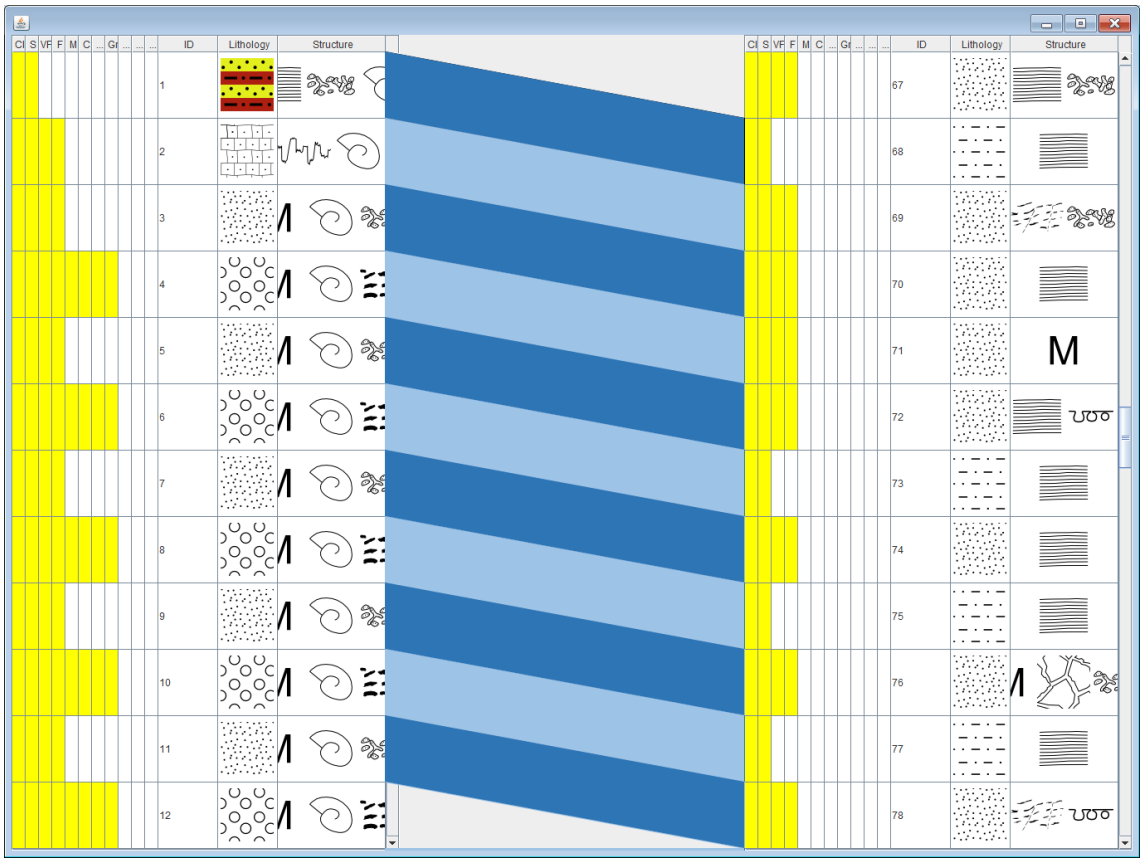

Figura 10.7: Trecho inicial de alinhamento gerado com o modelo de maior número de *clusters* e os parâmetros: *gap* = *6*; *match* = 3; *mismatch* = -3.

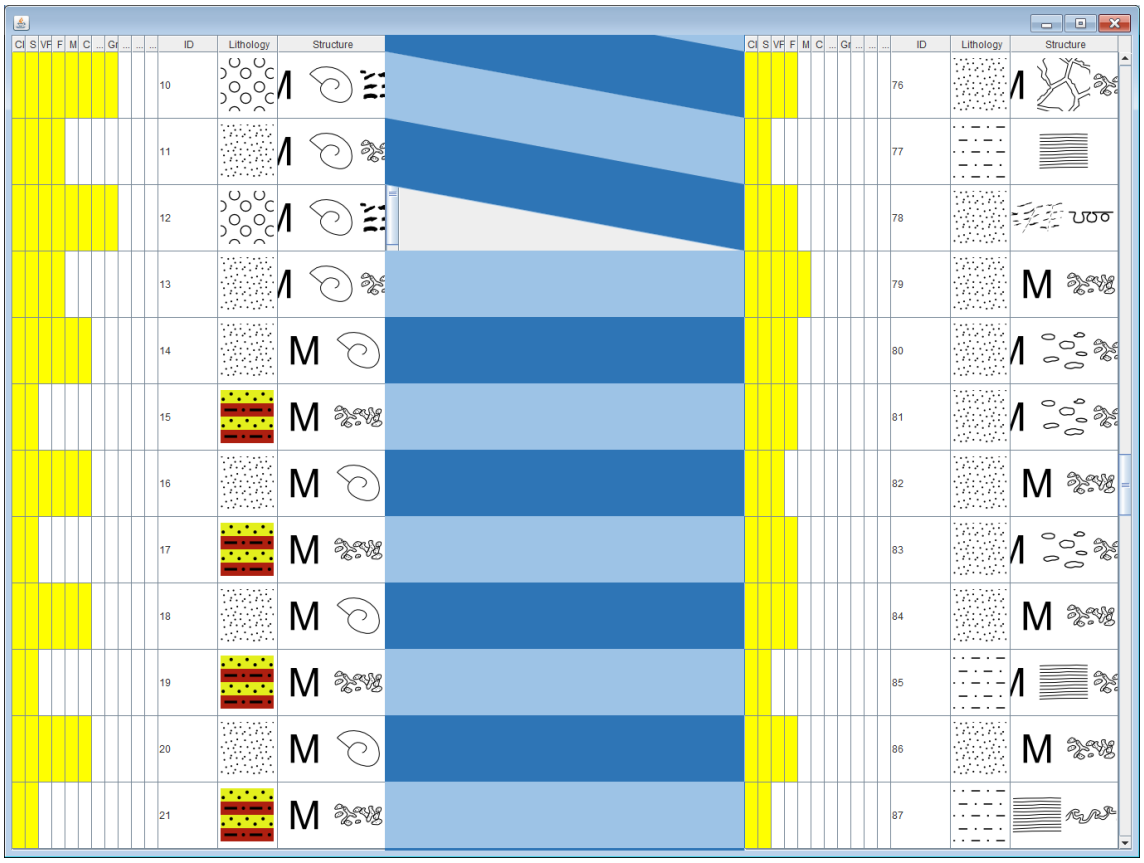

Figura 10.8: Trecho de alinhamento gerado com o modelo de maior número de *clusters* e os parâmetros: *gap* = *6*; *match* = 3; *mismatch* = -3.

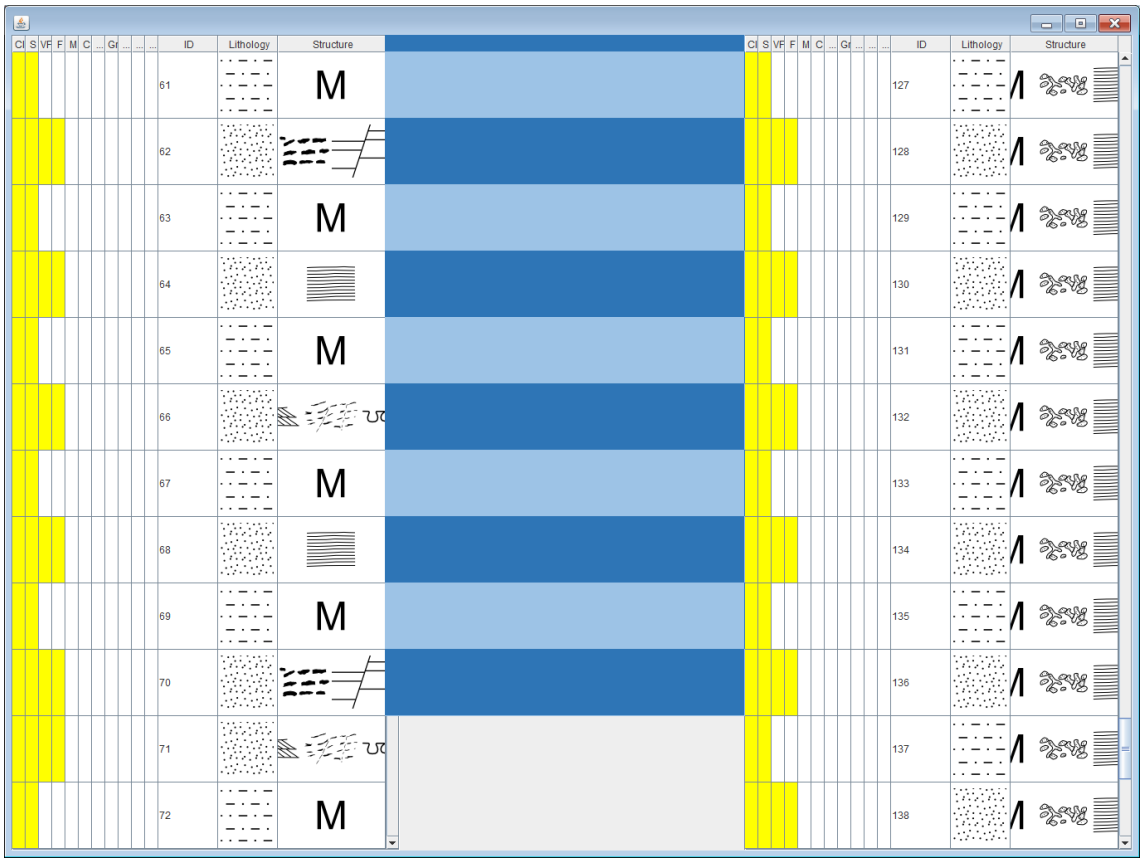

Figura 10.9: Trecho de alinhamento gerado com o modelo de maior número de *clusters* e os parâmetros: *gap* = *6*; *match* = 3; *mismatch* = -3.

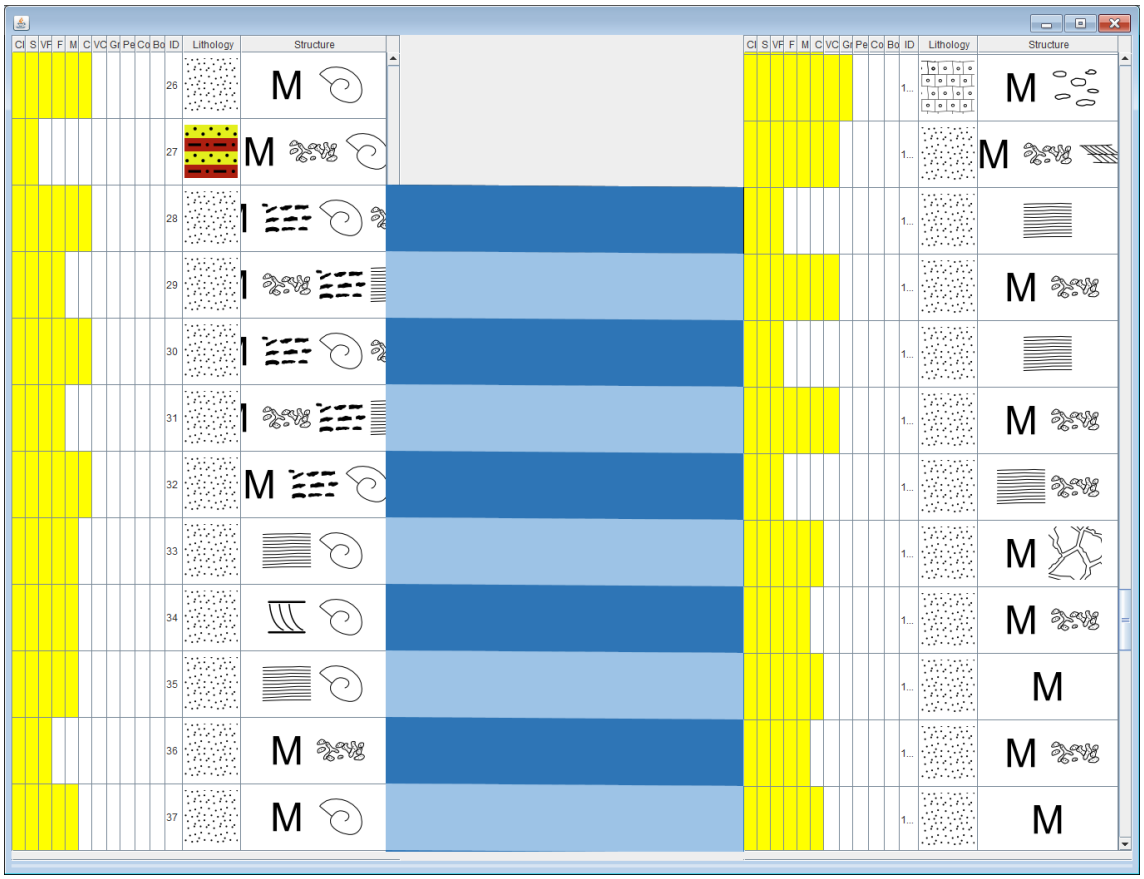

Figura 10.10: Trecho inicial de alinhamento gerado com o modelo de maior número de *clusters* e os parâmetros: *gap* = *13*; *match* = 3; *mismatch* = -10.

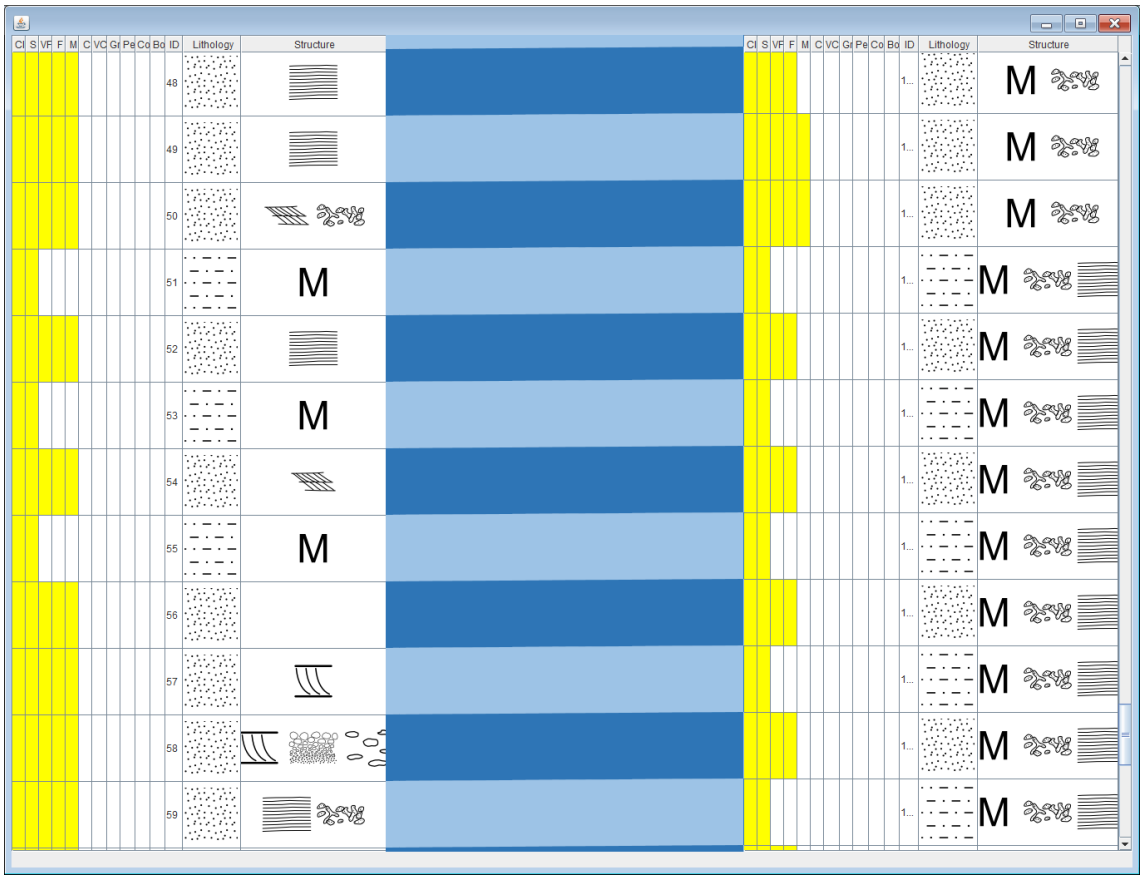

Figura 10.11: Trecho inicial de alinhamento gerado com o modelo de maior número de *clusters* e os parâmetros: *gap* = *13*; *match* = 3; *mismatch* = -10.

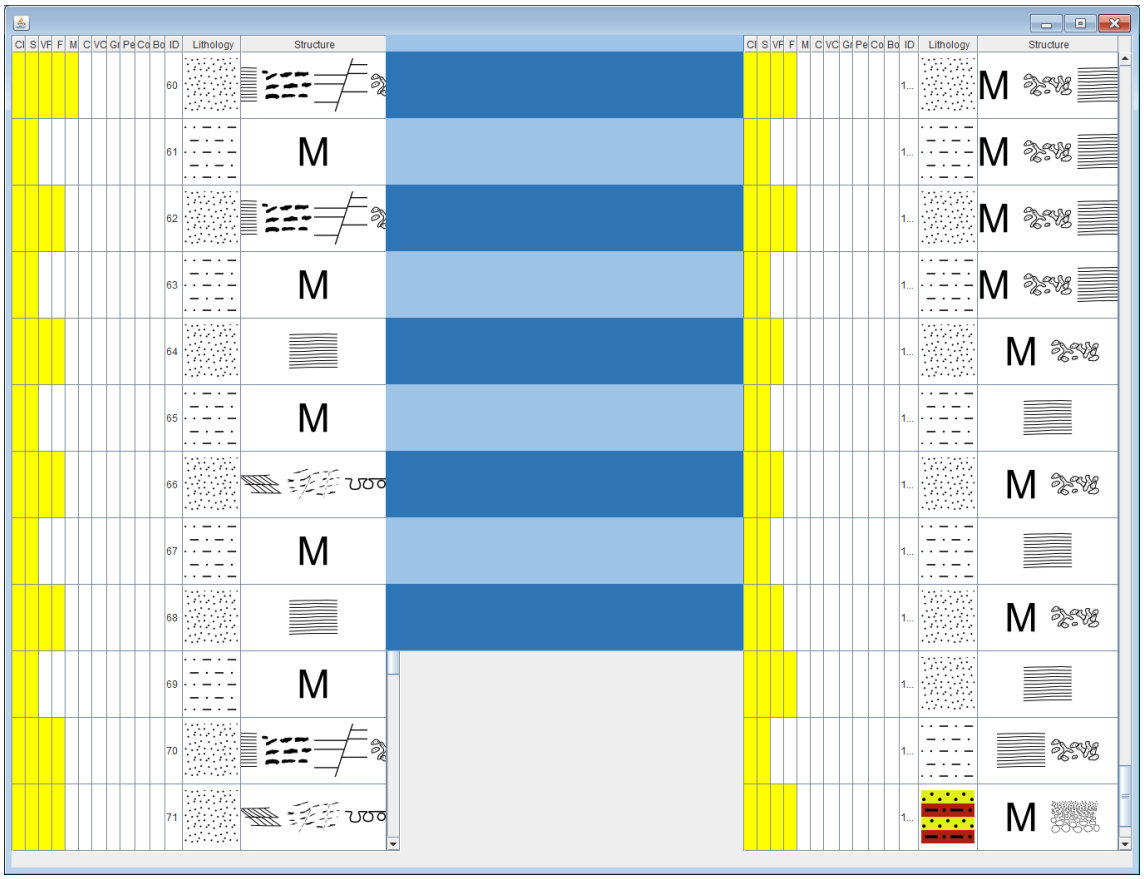

Figura 10.12: Trecho inicial de alinhamento gerado com o modelo de maior número de *clusters* e os parâmetros: *gap* = *13*; *match* = 3; *mismatch* = -10.

11 CONCLUSÕES

Neste trabalho, propomos uma abordagem para correlações litológicas automáticas a partir de descrições de testemunhos baseadas em ontologias. Um dos resultados é uma plataforma de pesquisa para gerar correlações que será utilizada para testar outras abordagens. O levantamento de dados resultou em *datasets* para testes de outras abordagens que podem vir a ser propostas no futuro. Além disto, diversas entrevistas foram realizadas e gravadas com geólogos neste trabalho, e permitirão estudos para desenvolvimento de trabalhos futuros.

Nos testes realizados, ficou evidente a importância do modelo de agrupamento de dados utilizado para gerar a correlação. A variação do modelo de agrupamento utilizado gera correlações bastante diferentes. Um problema é que estes modelos não podem ser gerados de maneira trivial. São necessários diversos testes e avaliações até que se encontre um modelo significativo. Outro problema é que os modelos são dependentes do *dataset* do qual foram gerados. Ou seja, não necessariamente os parâmetro utilizados para gerar um bom modelo serão suficientes para gerar um outro modelo a partir de um outro *dataset*.

Os algoritmos de agrupamento utilizados até o momento neste trabalho não consideram diferentes pesos de significância entre os atributos utilizados para descrever os objetos de domínio. Porém, no domínio da Geologia, não só alguns atributos possuem maior relevância na distinção de fácies sedimentares, como a importância dos atributos variam em função do ambiente geológico com que se está trabalhando. Trabalhos futuros também poderão investigar abordagens de agrupamento de dados que considerem a variação de pesos dos atributos. Além disso, algoritmos de agrupamento de dados não consideram uma gradação no nível de similaridade. Sua comparação é binária (pertence ou não pertence), enquanto neste domínio, temos um nível de similaridade entre fácies que não é binário, mas sim gradacional.

Constatou-se que os geólogos realizam correlações partindo de uma camada guia, chamada de *datum*, que possui características singulares dentro do poço e que aparece ao longo dos diversos poços que serão correlacionados. Ela é importante porque representa uma correspondência entre os poços que necessariamente deve ser mantida. No entanto, atualmente a abordagem proposta não considera esta noção. Isto sugere a necessidade de se investigar abordagens de alinhamento (talvez modificações dos algoritmos clássicos) que sejam capazes de realizar o alinhamento considerando subsequências previamente alinhadas como entrada. Isto será investigado em trabalhos futuros.

Verificou-se que geralmente os geólogos realizam correlações combinando descrições de testemunhos, como as utilizadas em nosso trabalho, com logs (numéricos) de petrofísica, como logs de raio gama, resistividade, etc. Deste modo, salienta-se a necessidade de se investigar no futuro abordagens que combinem ambos os tipos de informação para realizar correlações.

Durante as entrevistas com os geólogos contatou-se também que atualmente a comunidade da Geologia tem se interessado cada vez mais pela correlação estratigráfica. A correlação estratigráfica difere da correlação litológica porque considera o tempo geológico. Ou seja, nesta tarefa, busca-se correlacionar unidades que tenham sido geradas no mesmo tempo geológico, nas mesmas condições ambientais, e não simplesmente que possuam atributos visuais superficiais semelhantes. Trabalhos futuros que investiguem abordagens para a correlação estratigráfica poderiam utilizar abordagens semelhantes a que foi apresentada aqui. No entanto, os depoimentos do geólogos sugerem que o agrupamento de dados aplicados sobre descrições dos atributos meramente visuais das fácies, não é uma abordagem apropriada neste caso. Geólogos não utilizam apenas as características visuais superficiais de fácies para realizarem este tipo de correlação, mas também a interpretação dos ambientes e processos que formaram as fácies sedimentares. Isto será explorado em trabalhos futuros, combinando o trabalho aqui apresentado com o trabalho realizado por Carbonera em (2012), que desenvolve uma abordagem automática para interpretação dos processos deposicionais responsáveis pela gênese de uma determinada fácies sedimentar.

REFERÊNCIAS

BORST, W. N. **Construction of Engineering Ontologies for knowledge sharing and reuse**. 1997. Tese (Doutorado em Ciência da Computação) — University of Twente, Enschede, The Netherlands.

CARBONERA, J. L. **Raciocínio sobre conhecimento visual**: um estudo em estratigrafia sedimentar. 2012. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal do Rio Grande do Sul (UFRGS).

CHAO, K.-M.; ZHANG, L. **Sequence comparison**: theory and methods. [S.l.]: Springer, 2009. v.7.

CHEN, X. et al. Well log correlation based on wavelet transform and knowledge. In: INTELLIGENT PROCESSING SYSTEMS, 1997. ICIPS'97. 1997 IEEE INTERNATIONAL CONFERENCE ON, 1997. **Anais...** [S.l.: s.n.], 1997. v.2, p.1217–1219.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. **Journal of the Royal Statistical Society. Series B (Methodological)**, [S.l.], p.1–38, 1977.

FIORINI, S. R. **S-Chart**: um arcabouço para interpretação visual de gráficos. 2009. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal do Rio Grande do Sul.

GAN, G.; MA, C.; WU, J. **Data clustering**: theory, algorithms, and applications. [S.l.]: Siam, 2007. v.20.

GARCIA, L. F.; CARBONERA, J. L.; ABEL, M. Ontologias Aplicadas ao Problema de Correlação Litológica no Domínio da Geologia do Petróleo. In: SEMINAR ON ONTOLOGY RESEARCH IN BRAZIL, 6., 2013. **Anais...** [S.l.: s.n.], 2013. p.203–208.

GLUYAS, J.; SWARBRICK, R. **Petroleum geoscience**. [S.l.]: Wiley. com, 2009.

GRESSLY, A. **Observations géologiques sur le Jura Soleurois**. [S.l.]: Allgemeine schweizerische Gesellschaft für die gesammten Naturwissenschaften, 1841.

GRIFFITHS, C.; BAKKE, S. Interwell matching using a combination of petrophysically derived numerical lithologies and gene-typing techniques. **Geological Society, London, Special Publications**, [S.l.], v.48, n.1, p.133–151, 1990.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge Acquisition**, [S.l.], v.5, p.199–220, 1993.

GUARINO, N. Formal Ontology and Information Systems. In: INTERNATIONAL CONFERENCE ON FORMAL ONTOLOGY AND INFORMATION SYSTEMS (FOIS), 1998, Trento, Italy. **Proceedings...** IOS Press, 1998. p.3–15.

GUIZZARDI, G. **Ontological Foundations for Structural Conceptual Models**. Enschede, The Netherlands: Universal Press, 2005. 410p. (CTIT PhD Thesis Series, v.05-74).

GUIZZARDI, G. et al. Towards Ontological Foundations for the Conceptual Modeling of Events. In: **Conceptual Modeling**. [S.l.]: Springer, 2013. p.327–341.

GUIZZARDI, G.; WAGNER, G. Using the Unified Foundational Ontology (UFO) as a Foundation for General Conceptual Modeling Languages. In: POLI, R. et al. (Ed.). **Theory and Application of Ontologies**. [S.l.]: Springer-Verlag, 2010.

GUIZZARDI, R. S. S. et al. Towards an Ontological Account of Agent-Oriented Goals. In: CHOREN, R. et al. (Ed.). **Software Engineering for Multi-Agent Systems V**: research issues and practical applications. [S.l.]: Springer-Verlag, 2007. p.148–164. (Lecture notes in computer science, v.5).

GUYON, I.; VON LUXBURG, U.; WILLIAMSON, R. C. Clustering: science or art. In: NIPS 2009 WORKSHOP ON CLUSTERING THEORY, 2009. **Anais...** [S.l.: s.n.], 2009.

HALL, M. et al. The WEKA data mining software: an update. **ACM SIGKDD Explorations Newsletter**, [S.l.], v.11, n.1, p.10–18, 2009.

HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. [S.l.]: Morgan kaufmann, 2006.

HOWELL, J. A FORTRAN 77 program for automatic stratigraphic correlation. **Computers & Geosciences**, [S.l.], v.9, n.3, p.311–327, 1983.

HYNE, N. **Non-technical Guide to Petroleum Geology, Exploration, Drilling, and Production**. [S.l.]: PennWell Books, 2001.

LIM, J.-S.; KANG, J.; KIM, J. Artificial intelligence approach for well-to-well log correlation. In: SPE INDIA OIL AND GAS CONFERENCE AND EXHIBITION, 1998. **Anais...** [S.l.: s.n.], 1998.

LORENZATTI, A. **Ontologia para Domínios Imagísticos**: combinando primitivas textuais e pictóricas. 2009. Dissertação (Mestrado em Ciência da Computação) — UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL.

LORENZATTI, A. et al. Ontology for Imagistic Domains: combining textual and pictorial primitives. In: ER WORKSHOPS, 2009. **Anais...** Springer, 2009. p.169–178. (Lecture Notes in Computer Science, v.5833).

MANDOIU, I.; ZELIKOVSKY, A. **Bioinformatics algorithms**: techniques and applications. [S.l.]: Wiley. com, 2008. v.3.

MOUNT, D. W. **Bioinformatics**: sequence and genome analysis. [S.l.: s.n.], 2004. v.2.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of molecular biology**, [S.l.], v.48, n.3, p.443–453, 1970.

OLEA, R. A. CORRELATOR 5.2 a program for interactive lithostratigraphic correlation of wireline logs. **Computers & geosciences**, [S.l.], v.30, n.6, p.561–567, 2004.

PARSONS, S. B. **Historical Geology**. Acesso em: 14 de Julho de 2013, Disponível em: <http://www.ocean.odu.edu/~spars001/geology_112/laboratory/session_04/handout.html>.

PRESS, F. et al. **Para entender aTerra. 4.ed**. [S.l.]: Bookman,, 2004.

READING, H. G. **Sedimentary environments**: processes, facies and stratigraphy. [S.l.]: Wiley. com, 2009.

RODRIGUES, A. D. **Caracterização faciológica e estratigráfica dos sistemas mistos, siliciclásticos-carbonáticos do grupo Barra Nova, campo de fazenda Santa Luzia, Bacia do Espírito Santo**. 2010.

SAMMUT, C.; WEBB, G. **Encyclopedia of Machine Learning**. [S.l.]: Springer, 2011. (Springer reference).

SELLEY, R. C. **Elements of petroleum geology**. [S.l.]: San Diego:. Academic Press,, 1998.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. **Journal of molecular biology**, [S.l.], v.147, n.1, p.195–197, 1981.

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge Engineering: principles and methods. **Data and Knowledge Engineering**, [S.l.], v.25, n.1-2, p.161–197, March 1998.

SUGUIO, K. **Geologia Sedimentar**. São Paulo: Edgar Bücher LTDA, 2003. 400p.

USGS. **United States Geological Survey Core Description**. Acesso em: 04 de Dezembro de 2013, Disponível em: <http://woodshole.er.usgs.gov/openfile/of02-002/htmldocs/mms84cd.htm>.

WATERMAN, M. S.; RAYMOND JR, R. The match game: new stratigraphic correlation algorithms. **Mathematical geology**, [S.l.], v.19, n.2, p.109–127, 1987.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining**: practical machine learning tools and techniques: practical machine learning tools and techniques. [S.l.]: Elsevier, 2011.

WU, X.; NYLAND, E. Automated stratigraphic interpretation of well-log data. **Geophysics**, [S.l.], v.52, n.12, p.1665–1676, 1987.