IntergenicDB: evolução de uma solução *web* para disponibilização de dados genômicos de regiões intergênicas de genomas bacterianos

IntergenicDB: evolution of a web solution for providing genomic data from intergenic regions of bacterial genomes

DOI:10.34117/bjdv7n9-418

Recebimento dos originais: 07/08/2021
Aceitação para publicação: 23/09/2021

Daniel L. Notari
Área do Conhecimento de Ciências Exatas e Engenharias – Universidade de Caxias do Sul, Rua Francisco Getúlio Vargas, 1130, CEP 95070-560 - Caxias do Sul – RS – Brazil
dlnotari@ucs.br

Jovani Dalzochio
Área do Conhecimento de Ciências Exatas e Engenharias – Universidade de Caxias do Sul, Rua Francisco Getúlio Vargas, 1130, CEP 95070-560 - Caxias do Sul – RS – Brazil
jovanidalzochio@gmail.com

Camila R. T. Andrade
Área do Conhecimento de Ciências Exatas e Engenharias – Universidade de Caxias do Sul, Rua Francisco Getúlio Vargas, 1130, CEP 95070-560 - Caxias do Sul – RS – Brazil
camila.rachel@gmail.com

Jórdan R. Rosa
Área do Conhecimento de Ciências Exatas e Engenharias – Universidade de Caxias do Sul, Rua Francisco Getúlio Vargas, 1130, CEP 95070-560 - Caxias do Sul – RS – Brazil
jordan.rs2006@gmail.com

Ester Baccega
Área do Conhecimento de Ciências Exatas e Engenharias – Universidade de Caxias do Sul, Rua Francisco Getúlio Vargas, 1130, CEP 95070-560 - Caxias do Sul – RS – Brazil
ebaccega@ucs.br

Scheila de Ávila e Silva
Instituto de Biotecnologia – Universidade de Caxias do Sul, Rua Francisco Getúlio Vargas, 1130, CEP 95070-560 - Caxias do Sul – RS – Brazil
sasilva6@ucs.br

RESUMO
Existem alguns bancos de dados relacionados às sequências regulatórias da expressão gênica. No entanto, existe uma lacuna quando se trata de sequências intergênicas de bactérias com informações genéticas associadas. Assim, o portal IntergenicDB, um repositório de público de sequências genômicas, vem ao encontro dessa necessidade. Ele permite aos pesquisadores consultar as informações sobre as sequências de regiões intergênicas e, bem como as funções biológicas associadas por meio de uma interface amigável. Este artigo tem como objetivo descrever os procedimentos computacionais

empregados para a criação do repositório, seu histórico, bem como a estrutura e características dos dados.

Palavras-chave: banco de dados biológico, dados genômicos, regiões intergênicasde bactérias, IntergenicDB.

ABSTRACT
There are some databases related to regulatory sequences of gene expression. However, there is a gap when it comes to bacterial intergenic sequences with associated genetic information. Thus, the IntergenicDB portal, a public repository of genomic sequences, address this subject. It allows researchers to query information about the sequences of intergenic regions as well as the associated biological functions through a user-friendly interface. This article aims to relate the computational procedures used to create the repository and its history. In addition, there is the description of structure and characteristics of the data.

Keywords: biological database, genomic data, bacterial intergenic regions, IntergenicDB.

1 INTRODUÇÃO

Os avanços tecnológicos (seja de *hardware* ou *software*) permitem o estudo do genoma a partir de outra perspectiva (Xia et al., 2020; Leite et al.,2021). Gradualmente, nos últimos anos, a Biologia tem utilizado as ferramentas proporcionadas pela Informática para a resolução de problemas [Marx 2013; Leite et al. 2021].

Um dos maiores desafios da era pós-genômica é a determinação de quando, quais os genes são “ligados” e “desligados”. A diferença entre duas espécies está mais relacionada com a transcrição de seus genes do que com a estrutura destes em si [Jung et al. 2020; Lehninger et al. 2013]. Assim, o estudo da regulação gênica contribui para a construção do conhecimento a respeito da funcionalidade dos genes em diferentes espécies, na questão da diferenciação celular em organismos multicelulares, na resposta celular frente às mudanças ambientais, entre outras questões [Engstrom e Pfleger, 2017].

Os elementos regulatórios da transcrição gênica encontram-se em regiões denominadas intergênicas, as quais consistem em um segmento de DNA não transcrito que contém as sequências responsáveis pelo processo de regulação de início e término da expressão dos genes.

Em organismos procariotos, como bactérias e outros organismos unicelulares, as sequências intergênicas podem estar relacionadas a um ou mais genes [Zaha et al. 2014]. As informações biológicas relacionadas a uma determinada sequência intergênica, como gene(s) associado(s), papel biológico do gene e outras informações, contribuem para a ampliação do conhecimento biológico dos elementos regulatórios. O *IntergenicDB* é um

banco de dados público desenvolvido para o estudo de sequências intergênicas [Notari et al. 2014] e foi modelado para armazenar informações sobre a estrutura e funcionalidade das sequências intergênicas de bactérias.

O portal *IntergenicDB* surgiu para atender a necessidade do grupo de Bioinformática da Universidade de Caxias do Sul (UCS). O desenvolvimento foi realizado por alunos dos cursos de Graduação em Ciência da Computação e Sistemas de Informação, em seus trabalhos de conclusão de curso. O primeiro trabalho foi desenvolvido por Molin (2009) e teve como objetivo realizar a modelagem do banco de dados. Na sequência, Davanzo (2010) realizou a primeira implementação do portal *web* do *IntergenicDB*, disponibilizando para a comunidade acadêmicas as primeiras consultas. Para atender a necessidade de atualizações na interface e outras funcionalidades administrativas, Picolotto (2012) implementou uma interface de administração para os pesquisadores envolvidos no projeto. De modo a automatizar a carga de dados, Dalzochio (2014) implementou um importador para que a atualização do *IntergenicDB* fosse realizada de forma automatizada e com maior frequência. Com a carga de dados automatizada, o volume de dados armazenados aumentou e foi necessária a migração do *SGBD MySQL* para o *SGBD PostgreSQL*. Essa atualização foi realizada por Andrade (2017). Adicionalmente a atualização do SGBD, Andrade (2017) também realizou a atualização tecnológica da interface *web* do *IntergenicDB*. A fim de otimizar computacionalmente as consultas do *IntergenicDB*, Rosa (2017) implementou um *data warehouse*. Por fim, Baccega (2020) realizou uma atualização nos dados e atualizações tecnológicas no importador previamente desenvolvido.

Este artigo descreve a evolução tecnológica e a metodologia utilizada para construir o banco de dados *IntergenicDB*, as características do conjunto de dados disponibilizado e as aplicações que fazem uso dos seus dados. O artigo está organizado com as seções de Materiais e Métodos, Resultados e Conclusões.

2 IMPLEMENTAÇÃO E HISTÓRICO DA BASE DE DADOS

Esta seção apresenta as fontes de dados utilizadas, o modelo conceitual e lógico do conjunto de dados, bem como a descrição do *software* que realizou a integração de informações necessárias para a criação do *IntergenicDB*.

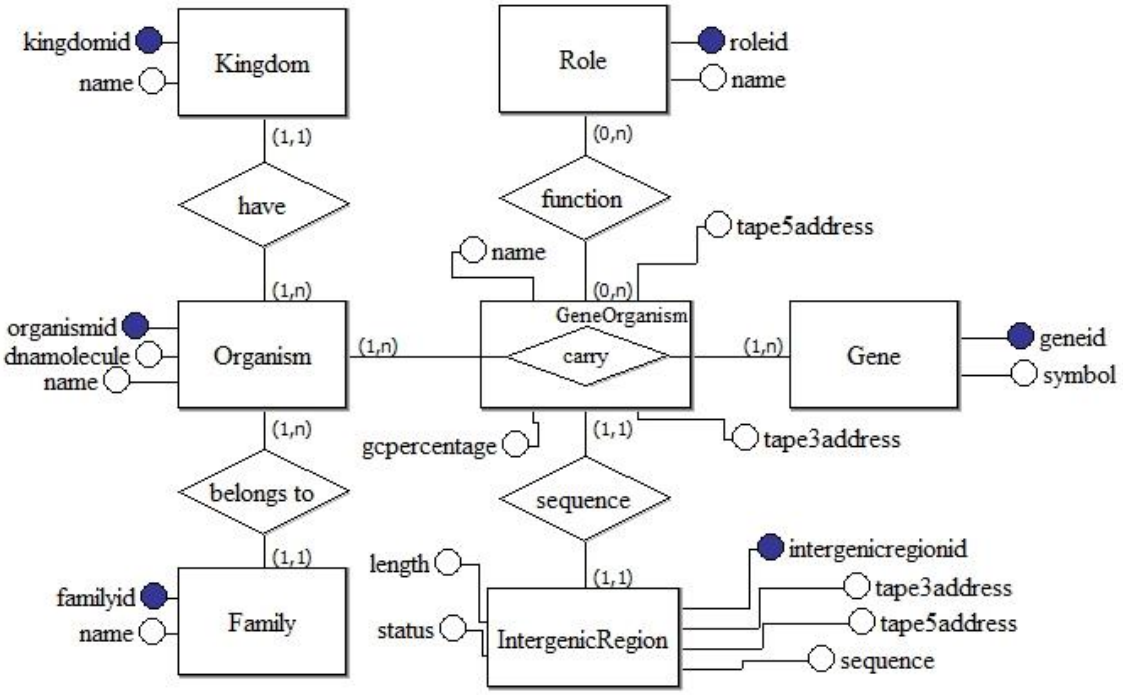
2.1 FONTE DE DADOS

Foram utilizados os dados dos bancos de dados biológico *GenBank* [SCHOCH et al., 2020] e *KEGG* [Kanehisa et al. 2017]. Para a coleta das informações relacionadas à sequência de nucleotídeos, como organismos, nome do gene, família e reino, foi realizado o *download* dos arquivos dos usando o *FTP* do *GenBank*. A única informação que não tem origem no *GenBank* é o papel biológico do gene. Para tal, foi realizada uma busca no banco de dados *KEGG* Brite utilizando a própria ferramenta disponível no seu *website*. Em relação ao histórico das fontes de dados do *IntergenicDB*, a versão inicial implementada por Davanzo (2010) contemplava as informações oriundas unicamente do *Genbank*. A incorpode informações oriundas do *KEGG* foi iniciada por Dalzochio (2014) e Rosa (2017).

2.2 MODELO DE DADOS

Em relação à modelagem de dados, a Figura 1 apresenta o diagrama Entidade-Relacionamento do banco de dados *IntergenicDB* utilizando a notação de Heuser (2009).

Figura 1: Modelo Conceitual do *IntergenicDB*



Fonte: Notari et *al.* (2017)

A relação entre os dados que incorporaram o *IntergenicDB* podem ser assim descrita [Notari et *al.*, 2017]: *(i)* cada região promotora está ligada a um organismo, que possui um nome, um reino, uma família, um tipo de molécula e um papel biológico; *(ii)*

cada gene possui um nome, um símbolo, um número de identificação de início e fim, uma função e uma quantidade percentual de CG; *(iii)* cada região intergênica possui um número para a posição inicial e final na sequência, um tamanho, sua sequência de nucleotídeos e o tipo de fita a que pertence.

A Figura 2 apresenta as tabelas que foram geradas a partir da aplicação e adaptação das regras de tradução de Heuser[1] (2009) no modelo conceitual da Figura 1.

Figura 2: Modelo Lógico do *IntergenicDB*

Family (familyid, name), Kingdom (kingdomid, name), Gene (geneid, symbol), Role (roleid, name)
Organism (organismid, name, dnamolecule, #familyid, #kingdomid)
GeneOrganism (geneorganismid, name, gcpercentage, tape5address, tape3address, #geneid, #organismid)
GeneOrganismRole (geneorganismroleid, #geneorganismid, #geneid)
IntergenicRegion (intergenicregionid, tape5address, tape3address, sequence, length, status, #aeneoranismid)

Fonte: Notari et *al.* (2017)

As tabelas "*Family"* e "*Kingdom"* possui os dados de todas as famílias e do reino dos organismos. A tabela "*Organism"* possui a informação da espécie, contemplando o nome científico completo. A tabela "*Gene"* contém os dados dos genes para cada organismo. A tabela "*Role"* possui os dados sobre que papel o gene desempenha. A tabela "*IntergenicRegion"* armazena a sequência de nucleotídeos das regiões intergênicas de cada gene. A tabela "*GeneOrganism"* contém os dados de cada gene de um organismo. E, por fim, a tabela "*GeneOrganismRole"* contém os dados dos papéis biológicos de cada gene de um organismo.

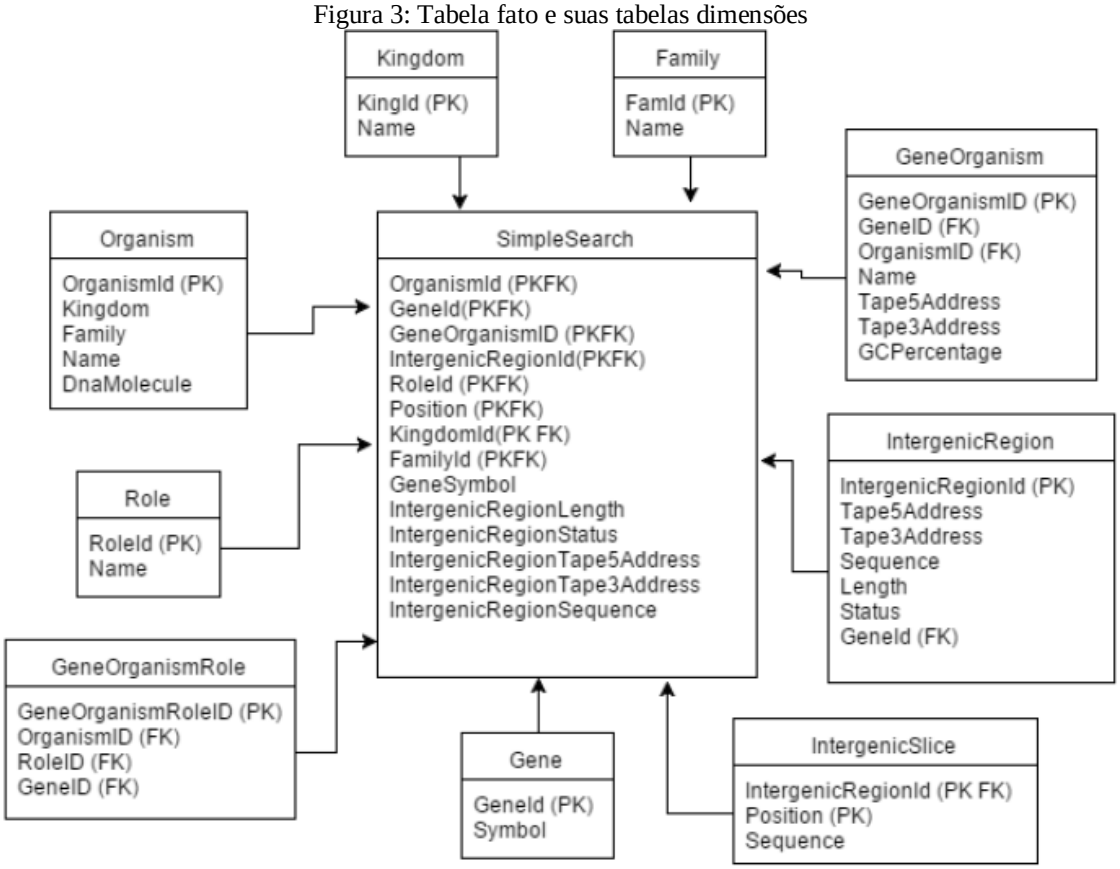
Em relação à evolução do modelo de dados, a primeira versão foi apresentada por Molin (2009), sendo essa modelagem aprimorada por Davanzo (2010), Dalzochio (2014)

[1] Coluna sublinhada indica chave primária, enquanto *hashtag* (#) indica chave estrangeira

e Rosa (2017). Todas as modificações foram incrementais e, muitas foram motivadas devido à necessidade de aprimoramento das consultas, objetivos dos pesquisadores e *feedback* dos usuários. Das modificações realizadas, a modificação com maior impacto foi a implementação do *data warehouse* utilizando o modelo denominado "Estrela", efetuada por Rosa (2017).

O objetivo da implementação de Rosa (2017) foi otimizar o desempenho da consulta à base dados, devido aos diversos relacionamentos e volume resultante. A tabela fato (Figura 3) foi formada somente pelos identificadores das tabelas dimensões. A tabela fato é estática, ou seja, os novos dados não são adicionados a tabela no momento da sua inserção. Para isso, é necessário rodar uma rotina que destrói a tabela fato e a recria.

Figura 3: Tabela fato e suas tabelas dimensões



Fonte: Rosa (2017)

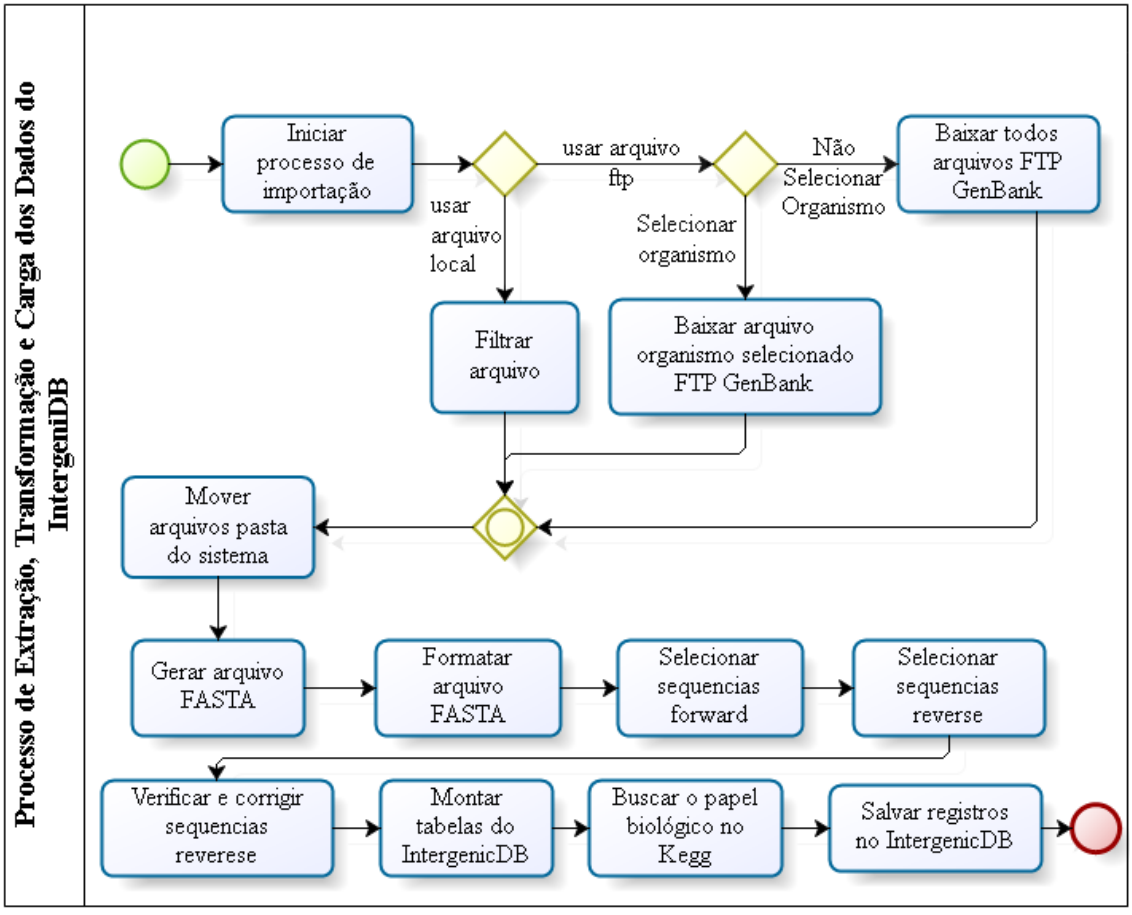
Adicionalmente à implementação do *data warehouse*, Andrade (2017) realizou a migração de tecnologia do *SGBD MySQL* para o *SGBD PostgreSQL* foi necessária devido a algumas limitações do *SGBD MySQL*, como: dificuldade na criação de *scripts SQL* compatíveis com o *MySQL*, para consultar os dados diretamente no servidor do banco de dados, lentidão nas consultas dos dados por meio da interface e erros na exportação dos

dados do banco por meio da ferramenta *phpMyAdmin*. Assim, para garantir o sucesso na migração, foram realizadas adequações nas estruturas das tabelas existentes, uma nova carga de dados foi realizada e, vários testes para garantir a integridade dos dados foram realizados.

2.3 IMPORTADOR DE DADOS

A agregação das informações foi realizada por meio da implementação de uma série de *scripts* executados na forma de um *workflow*. Todos os *scripts* do importador foram desenvolvidos em linguagem *Python* e a interface do importador foi desenvolvida em C#. A esta implementação designou-se o nome de *MMDBImportTool* [Dalzochio 2014]. Nela é possível configurar o diretório *FTP* do *GenBank* para a busca dos dados, bem como, buscar as informações banco de dados *KEGG BRITE*. A Figura 4 apresenta a sequência de atividades que foram realizadas para a geração dos dados para o *IntergenicDB*.

Figura 4: Atividades executadas pelo Importador de Dados

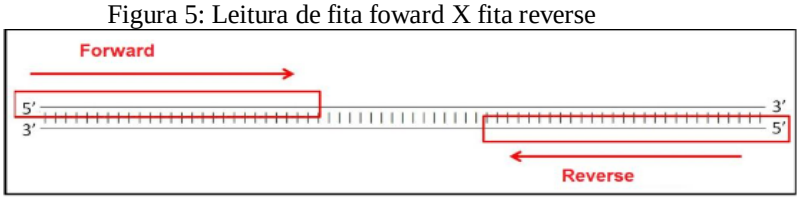


Fonte: Dalzochio (2014)

O processo foi iniciado com a configuração do acesso ao *FTP* do *GenBank* com a seleção dos arquivos dos organismos a serem processados. Estes arquivos foram baixados e salvos em uma pasta local para o posterior processamento dos dados. Na sequência, um arquivo do tipo *FASTA* foi gerado para ser utilizado pelos outros *scripts*. Este arquivo foi utilizado na etapa seguinte, que consistiu na separação dos dados da fita dupla de DNA. A separação foi realizada de acordo com a posição da sequência da fita de DNA: *forward* e *reverse* [De Robertis, 2017]. As informações *foward* e *reverse* devem ser tratadas de forma diferente, pois como pode ser visto na Figura 5, a informação da posição de início da sequência intergênica muda conforme sua localização no DNA.

As sequências reversas foram reescritas, invertendo-as e depois substituindo os nucleotídeos conforme o dogma da Biologia Molecular [De Robertis, 2017]. O penúltimo passo envolveu buscar as informações do papel biológico no repositório *KEGG*. Por fim, os dados foram salvos na tabela do banco de dados. Uma vez que esse processo foi finalizado, as informações foram armazenadas no banco de dados *IntergenicDB*.

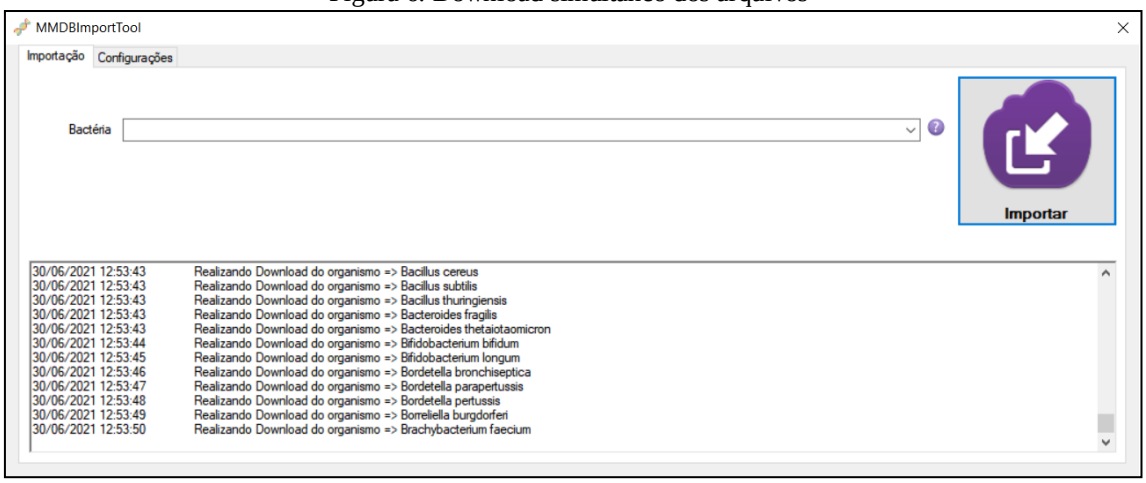
Figura 5: Leitura de fita foward X fita reverse



Fonte: Notari et *al.* (2017)

Em 2020 o importador foi reavaliado e aprimorado. Para isso, foram atualizadas bibliotecas, testagem dos acessos externos utilizados e foi implementado o uso de *Threads* [Baccega 2020].

Figura 6. Download simultâneo dos arquivos



Fonte: Baccega (2020)

A paralelização do *MMDBImportTool* utilizou a classe *System.Threading.Tasks* da própria linguagem *C#*. Desta forma, foi realizada a utilização das *Tasks* (tarefas concorrentes) para o *download* dos arquivos, para extração dos dados dos arquivos *FASTA* no sentido *forward* e no sentido *reverse*. Esse processo possibilitou que o *download* de múltiplos arquivos fossem realizados de forma simultânea (Figura 6). Ao término do *download* de cada arquivo, uma cópia foi adicionada a um diretório criado automaticamente pela ferramenta de importação. Na sequência, foi inicializada a execução dos *scripts* na linguagem *Python* para geração dos arquivos no formato *FASTA* e então, também de forma simultânea com a utilização de *Tasks*, foi iniciada a extração dos dados desses arquivos, nos sentidos *forward* e *reverse*.

3 RESULTADOS: CONJUNTO DE SEQUÊNCIAS E INTERFACE DE DISPONIBILIZAÇÃO

A execução dos passos do Importador *MMDBImportTool* [DALZOCHIO 2014] descritos na Figura 3 geraram os dados salvos nas tabelas do *IntergenicDB* descritas na Figura 2. O Quadro 1 apresenta a relação entre os dados importados do GenBank e os dados armazenados no *IntergenicDB*. Os dados do Quadro 1 compreendem: *(i)* Origem:

informa o dado extraído do arquivo de dados; (*ii)* Destino: descreve a tabela e coluna na qual o dado foi inserido; e *(iii)* Descrição: identifica o conteúdo do dado inserido.

Quadro 1: Mapeamento entre os dados extraídos do *GenBank* com as tabelas do *IntergenicDB*

Origem	Destino	Descrição
Genbank (tag organism)	*Family.Name*	Nome da família do organismo
Genbank (tag organismo	*Kingdom.Name*	Reino do organismo
Genbank (tag organism)	*Organism.Name*	Nome do organismo
Genbank (tag mol_type)	*Organism.DnaMolecule*	Molécula de DNA
Genbank (tag gene)	*Gene.Symbol*	Símbolo do gene
KEGG	*Role.Name*	Papel biológico do gene
Genbank (tag product)	*GeneOrganism.Name*	Nome do organismo
Genbank (gerado scripts)	*GeneOrganism.GCPercentage*	% de GC da sequência de DNA
Genbank (tag complement)	*GeneOrganism.Tape3Address*	Fita de DNA sentido *forward*
Genbank (complement)	*GeneOrganism.Tape5Address*	Fita de DNA sentido *reverse*
Genbank (usa arquivo FASTA)	*IntergenicRegion.Sequence*	Sequência região intergênica
Genbank (usa arquivo FASTA)	*IntergenicRegion.Tape3Address*	Posição sequência *forward*
Genbank (usa arquivo FASTA)	*IntergenicRegion.Tape5Address*	Posição sequência *reverse*
Genbank (usa arquivo FASTA)	*Length*	Tamanho da região intergênica
Genbank (Genbank)	*Status*	F – *Foward*, R – *Reverse*

As seções a seguir apresentam os detalhes de testes realizados e a comparação entre as duas cargas de dados realizadas [Rosa 2017] [Baccega 2020].

3.1 IMPORTAÇÃO DOS DADOS

A importação dos dados foi realizada em três etapas. Inicialmente foram realizadas importações de organismos individuais, para verificar se os dados gravados nas tabelas estavam consistentes com o arquivo importado. Para essa fase, cerca de 15 organismos foram escolhidos de forma aleatória para a importação. Durante a primeira fase, nenhum dos organismos apresentou erros, sendo que todas suas informações importadas com sucesso.

Na segunda fase, a ferramenta de importação foi executada de forma completa e automatizada, com seleção de todos os organismos disponibilizados pelo *Genbank*.

Quando a importação foi concluída, foram observadas inconsistências nos dados. Durante a execução da ferramenta, em determinados momentos ocorreram erros de *timeout* do banco de dados. Além disso, por alguma razão não conhecida – podendo ter como causas o próprio ambiente de testes, a *MMDBImportTool* parava sua execução, porém não apresentava nenhum tipo de erro em tela. Desta forma, alguns arquivos de dados não eram processados até o final, e com isso, não eram gravadas todas as informações no banco de dados. Como os erros descritos não apresentavam nenhum padrão, foi realizada a terceira fase da importação.

A terceira fase foi iniciada em um banco de dados novo e durante essa fase, os organismos foram importados de forma manual, até que todos fossem importados. Com a importação de forma controlada, os organismos eram importados um a um, realizando anotações sobre cada um deles. Então, foi possível identificar um padrão para os erros que estavam acontecendo. Para alguns organismos, a importação ocorria apenas no sentido *forward*, ocorrendo erro ao tentar executar a extração de dados no sentido *reverse*. Cerca de 20 organismos se encontravam nessa situação e foram marcados para novos testes após o término da importação.

Ainda durante a terceira fase, juntamente com os erros no sentido *reverse*, foram identificados arquivos do *Genbank* de organismos que não tinham a cadeia de caracteres de entrada correta. Foi realizada uma verificação em todos os arquivos desses organismos, mas novamente, nada anormal foi encontrado. Para esses organismos, foi realizada uma nova solicitação de *download*, para buscar por possíveis correções fornecidas pelo *Genbank* e realizar uma nova tentativa para a importação, mas o erro permaneceu. Entre esses organismos com cadeira incorreta então *Bordetella pertussis Tohama*, *Buchnera aphidicola*, *Escherichia coli IAI39*, *Shigella dysenteriae* e *Streptococcus agalactiae*.

3.2 CARGA DE DADOS

A carga de dados do *IntergenicDB* realizada por Rosa (2017) apresentou 15095 genes, 17 famílias e 88 organismos. A nova carga de dados, realizada por Baccega (2020), consistiu em duas formas: uma carga automática e uma carga individual para cada organismo. A primeira carga completa (carga automatizada) realizada com o novo importador gerou 7042 genes, 17 famílias e 110 organismos. A carga final, realizada de forma individualizada, gerou 9209 genes, 18 famílias e 118 organismos. Esses dados são apresentados na Tabela 1.

Tabela 1 – Comparativo da carga de dados

	Carga de Dados Rosa (2017)	Carga Automatizada Baccega (2017)	Carga Individualizada Baccega (2017)
Genes	15095	7043	9209
Famílias	17	17	18
Organismo	88	110	118

Apesar da carga final (realizada de forma individualizada) conter mais organismos, ainda assim, possui menos genes registrados (diferença de 8114 genes) que a carga de Rosa (2017). No entanto, houve um aumento em relação a primeira carga completa. É possível justificar essa diferença de registros como uma consequência dos problemas de importação da carga citados na seção 3.1. Em relação a quantidade de organismos, na primeira carga, o número aumentou em 22 organismos na comparação com a carga realizada por Rosa (2017), porém, na carga final, esse número aumentou em 30 novos organismos importados, totalizando 118 registros. O número de registros de famílias, se manteve igual na comparação da carga de dados de Rosa (2017), com a carga automatizada, porém em comparação com a carga final, aumentou em 1 registro, passando de 17 para 18 famílias.

3.3 ANÁLISE DO PARALELISMO

Para análise dos resultados foram feitas medidas de tempo considerando a aplicação sequencial e a aplicação que efetua o *download* simultâneo utilizando *threads*. A ferramenta *MMDBImportTool* permite realizar a importação de apenas um organismo específico ou de todos os 109 organismos disponibilizados pelo *Genbank*. Desta forma, foram realizados testes comparativos realizando o *download* de todos os organismos e utilizando 3 organismos escolhidos de forma aleatória: *Bacillus anthracis*, *Escherichia coli* e *Klebsiella pneumoniae*. Na Tabela 2 tem-se um comparativo dos resultados. Nela são apresentados (em segundos): *(i)* o tempo total de execução¸ *(ii)* o tempo necessário somente para o download dos arquivos, e *(iii)* o tempo necessário somente para a extração dos dados dos arquivos.

Tabela 2. Tempo de Execução (em segundos)

	ANTES			**DEPOIS**		
	Download	Extração	Total de processamento	*Download*	Extração	Total de processamento
Bacillus anthracis	41	266	307	42	228	570
Escherichia coli	38	632	670	38	523	561
Klebsiella pneumoniae	76	21	97	70	18	88

Carga completa	8700	27780	36480	7860	22620	30480

Como pode ser observado, apesar das alterações realizadas no processo de *download*, não houve mudanças no tempo de execução deste procedimento. Neste caso, o limitador é a rede e a velocidade do *download*, uma vez que a banda é dividida entre os *downloads* que são realizados simultaneamente. Em relação a extração das informações e processamento dos arquivos, observa-se uma diminuição no tempo de execução, tanto para os organismos específicos como para o processamento dos 109 organismos.

3.4 CONJUNTO DE DADOS

O conjunto de dados gerado no formato CSV para pesquisadores utilizarem envolveu a execução de uma consulta em linguagem SQL no banco de dados IntergenicDB, conforme mostrada na Figura 7

Figura 7. Consulta SQL para gerar o DataSet

```
Select o.name as organism, g.symbol as gene, k.name as kingdom, f.name as family, r.name as role,
ir.tape5address, ir.tape3address, ir.sequence
from organism o, gene g, geneorganism go, kingdom k, family f, role r,geneorganismrole gor,
intergenicregion ir
where o.familyid = f.familyid and o.kingdomid = k.kingdomid and o.organismid = go.organismid and
g.geneid = go.geneid and go.geneorganismid = gor.geneorganismid and r.roleid = gor.roleid and
go.geneorganismid = ir.geneorganismid
```

A estrutura do *dataset* gerado pela execução da consulta da Figura 7 possui a seguinte estrutura:

- *organism*: nome do organismo
- *gene*: símbolo do gene
- *kingdom*: reino do organismo
- *family*: nome da família do organismo
- *role*: papel biológico do gene
- *tape5address*: posição da sequência no sentido *forward*
- *tape3address*: posição da sequência no sentido *reverse*
- *sequence*: sequência de dados da região intergênica

3.5 INTERFACE *WEB*

O objetivo do *IntergenicDB* é ser um portal de consultas para regiões intergênicas sobre os dados armazenados de organismos procariontes [Avila e Silva 2011; Notari 2012]. O portal possui uma área administrativa de acesso privado e uma área de consulta pública.

Em relação à área de consulta pública, a primeira versão da interface *web* foi implementada por Davanzo (2010). Neste trabalho, o objetivo principal foi a criação de uma interface de acesso ao banco de dados *IntergenicDB*, realizando consultas e inserção de dados. A interface *web* foi migrada para plataforma de desenvolvimento *PHP*. Esta alteração proporcionou uma evolução no *layout* do portal e a inclusão de novas funcionalidades (Andrade, 2017).

Em relação à apresentação dos resultados das pesquisas, a consulta disponibilizada no portal do *IntergenicDB*, possui interface baseada em *query builders* dos portais como, por exemplo, *"PubMed Advanced Search Builder"* do NCBI[2], "*Free text search"* do EBI[3] "*ARSA"* do DDBJ[4] e outros. A partir dos parâmetros selecionados pelo usuário, o mecanismo de consulta monta uma consulta SQL. O usuário necessita possuir conhecimento em operadores lógicos para montar a pesquisa que deseja [Rosa 2017]. Uma vez com os parâmetros informados, o mecanismo gera a consulta buscando os identificadores dos parâmetros em nas tabelas dimensões e assim, após obtê-las, a consulta segue na tabela fato. Com os identificadores filtrados a partir da tabela fato, a consulta busca seus nomes nas tabelas dimensões. A consulta somente ocorre no dentro do intervalo de linhas da paginação determinada pelo usuário, ou seja, se o usuário parametrizar na consulta que deseja visualizar 50 linhas por página, a consulta somente ocorrerá dentro desse intervalo para minimizar o volume de dados a ser manipulado.

[2] https://www.ncbi.nlm.nih.gov/pubmed/advanced
[3] http://www.ebi.ac.uk/ena/data/warehouse/search
[4] http://ddbj.nig.ac.jp/arsa/

Em relação à área administrativa do portal pode ser acessada mediante usuário e senha. O acesso a essa área possibilita o gerenciamento dos cadastros de usuários e grupos do portal, sendo que o grupo ao qual o usuário está associado lhe proverá mais ou menos acessos. É possível também consultar individualmente os dados biológicos inseridos no banco de dados, não sendo permitida a inserção, nem a manutenção dos dados. A área administrativa ainda provê aos administradores acesso a dados de geoposição dos usuários que acessam o portal e a administração dos textos das páginas de Início e Ajuda do portal [Picolotto 2012].

4 Conclusão

O desenvolvimento de um banco de dados para integrar os dados das regiões intergênicas em um repositório único mostra-se uma ferramenta auxiliar ao pesquisador que requer este tipo de informação. Uma vez que os dados de genomas de bactérias são constantemente atualizados.

O presente artigo descreveu a evolução do repositório de dados *IntergenicDB*. Esse repositório tem capacidade de tornar-se um grande portal pois, em relação aos demais portais estudados, possui um tratamento único sobre as informações relacionadas a regiões intergênicas.

Adicionalmente, está em desenvolvimento a conexão entre as sequências depositadas neste banco de dados com outras ferramentas on-line de análise de regiões regulatórias. Considerando esta questão, este projeto visa integrar as regiões intergênicas armazenadas no IntergenicDB com outras ferramentas disponíveis na internet para análise de dados e/ou predição de sequências.

REFERÊNCIAS

Andrade, C. R. T. D. (2017) “IntergenicDB 2.0”. Trabalho de Conclusão do Curso de Sistemas de Informação da Universidade de Caxias do Sul.

Avila e Silva, S. (2011) “Redes neurais artificiais aplicadas no reconhecimento de regiões promotoras em bactérias Gram-negativas”. Tese de Doutorado, Programa de Pós-Graduação em Biotecnologia da Universidade de Caxias do Sul.

Baccega, E. (2020). “Evolução do Intergenicdb: banco de dados de regiões intergênicas de bactérias gram-negativas”. Trabalho de Conclusão do Curso de Ciência da Computação da Universidade de Caxias do Sul.

DAVANZO, Vanessa. Desenvolvimento de Consultas para um Banco de Dados de Sequências Intergênicas. 2010. 76 f. TCC (Graduação) - Curso de Bacharelado em Ciência da Computação, Universidade de Caxias do Sul, Caxias do Sul, 2010. Computação, Universidade de Caxias do Sul, Caxias do Sul, 2014.

Dalzochio, J. (2014) “Implementação de novas funcionalidades para o portal IntergenicDB e povoamento do seu banco de dados”. Trabalho de Conclusão do Curso de Sistemas de Informação da Universidade de Caxias do Sul.

De Robertis, E. D. P. (2017) Bases da biologia celular e molecular. 16. ed. Rio de Janeiro: Grupo Gen.

ENGSTROM, Michael D.; PFLEGER, Brian F. Transcription control engineering and applications in synthetic biology. Synthetic And Systems Biotechnology, [S.L.], v. 2, n. 3, p. 176-191, set. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.synbio.2017.09.003.
Gannon, D. & Reed, D. (2009) “Parallelism and the cloud, The fourth paradigm: Data intensive scientific discovery”. T. Hey et al., eds. Washingtong: Microsoft Research, 131-135.

Heuser, C. (2009) Projeto de Banco de Dados. Porto Alegre: Bookman, v. 4.

JUNG, Hyungtaek; VENTURA, Tomer; CHUNG, J. Sook; KIM, Woo-Jin; NAM, Bo-Hye; KONG, Hee Jeong; KIM, Young-Ok; JEON, Min-Seung; EYUN, Seong-Il. Twelve quick steps for genome assembly and annotation in the classroom. Plos Computational Biology, [S.L.], v. 16, n. 11, p. 1008325e, 12 nov. 2020. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pcbi.1008325

Kanhere, A.; Bansal, M. (2005) “Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes”. Nucleic Acids Research 33:3165-3175.
Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017) “KEGG: new perspectives on genomes, pathways, diseases and drugs”. Nucleic Acids Research, 45(Database issue), D353–D361.

LEITE, Michel L.; COSTA, Lorena S. de Loiola; CUNHA, Victor A.; KRENISKI, Victor; BRAGA FILHO, Mario de Oliveira; CUNHA, Nicolau B. da; COSTA, Fabricio F. (2021) Artificial intelligence and the future of life sciences. Drug Discovery Today, [S.L.], p. 1, jul. 2021. Elsevier BV. http://dx.doi.org/10.1016/j.drudis.2021.07.002

Lehninger, A. L.; Cox, M. M.; Nelson, D. L. (2013) Principles of biochemistry. 6. ed. New York: Worth.

Marx, V. (2013) “Biology: The big challenges of big data”. Nature 498, 255–260.

MOLIN, Aurione Francisco. Prototipagem de um Banco de Dados de Promotores como base para um Portal de Serviços. 2009. 74 f. TCC (Graduação) - Curso de Bacharelado em Ciência da Computação, Universidade de Caxias do Sul, Caxias do Sul, 2009.

Naghdi, M. R., Smail, K., Wang, J. X., Wade, F, Breaker, R. R, Perreault, J. (2017) “Search for 5′-leader regulatory RNA structures based on gene annotation aided by the RiboGap database”, Methods, Vol. 117, Pages 3-13.

Notari, D. L. (2012) “Desenvolvimento de workflows científicos para a geração e análise de diferentes redes de interatomas”. Tese de Doutorado, Programa de Pós-graduação em Biotecnologia, Universidade de Caxias do Sul.

Notari, D. L., Molin, A., Davanzo, V., Picolotto, D., Ribeiro, H. G., & Silva, S. de A. e. (2014) “IntergenicDB: a database for intergenic sequences”. Bioinformation, 10(6), 381–383.

Notari DL; Dalzochio, J. ; Andrade, C. R. T. ; Rosa, J. R. ; Klauck, H. A. ; Silva, S. A. . IntergenicDB: Banco de dados de regiões intergênicas de Bactérias Gram-Negativas. In: Brazilian Symposium on Databases, 2017, Uberlância. SBBD Proceedings Satellite Events of the 32nd Brazilian Symposium on Databases. Uberlandia: online, 2017. v. 1. p. 234-244.

Pearson, W. and Lipman, D. (1988) “Improved tools for biological sequence comparison (amino acid/nucleic acid/data base searches/local similarity)”. Proceedings of the National Academy of Sciences, 85, 2444-2448.

PICOLOTTO, Douglas. Implementação de novas funcionalidades para o portal IntergenicDB. 2012. 99 f. TCC (Graduação) - Curso de Curso de Bacharelado em Sistemas de Informação, Universidade de Caxias do Sul, Caxias do Sul, 2012.

Rosa, J. (2017) “Criação e implementação de um data warehouse para reformatar o mecanismo de pesquisa do portal IntergenicDB 2.0”. Trabalho de Conclusão do Curso de Ciência da Computação da Universidade de Caxias do Sul.

SCHOCH, Conrad L; CIUFO, Stacy; DOMRACHEV, Mikhail; HOTTON, Carol L; KANNAN, Sivakumar; KHOVANSKAYA, Rogneda; LEIPE, Detlef; MCVEIGH, Richard; O’NEILL, Kathleen; ROBBERTSE, Barbara. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database, [S.L.], v. 2020, p. 1, 1 jan. 2020. Oxford University Press (OUP). http://dx.doi.org/10.1093/database/baaa062.

Scott, E. Y., Mansour, T., Bellone, R. R., Brown, C. T., Mienaltowski, M. J., Penedo, M. C., … Finno, C. J. (2017). “Identification of long non-coding RNA in the horse transcriptome”. BMC Genomics, 18, 511.

Tong, C., Chen, Q., Zhao, L., Ma, J., Ibeagha-Awemu, E. M., & Zhao, X. (2017). "Identification and characterization of long intergenic noncoding RNAs in bovine mammary glands". BMC Genomics, 18, 468.

XIA, Jianyang; WANG, Jing; NIU, Shuli. Research challenges and opportunities for using big data in global change biology. Global Change Biology, [S.L.], v. 26, n. 11, p. 6040-6061, 13 set. 2020. Wiley. http://dx.doi.org/10.1111/gcb.15317

Zaha, A; Ferreira, H. B.; Passaglia, L. M. P. (2014) Biologia Molecular Básica. Porto Alegre: Artmed.