Fundação Educacional do Município de Assis
Instituto Municípal de Ensino Superior de Assis - IMESA

RENATO APARECIDO FRIZANCO

UM ESTUDO SOBRE A BIOINFORMÁTICA: DISTINÇÃO DOS TIPOS DE DADOS E AS DIFERENTES BASES DE DADOS EXISTENTES

ASSIS
2010

Fundação Educacional do Município de Assis
Instituto Municípal de Ensino Superior de Assis - IMESA

RENATO APARECIDO FRIZANCO

UM ESTUDO SOBRE A BIOINFORMÁTICA: DISTINÇÃO DOS TIPOS DE DADOS E AS DIFERENTES BASES DE DADOS EXISTENTES

Trabalho de Conclusão de Curso apresentado ao Instituto Municipal de Ensino Superior de Assis, como requisito de Curso de Graduação, analisado pela seguinte comissão examinadora:

Orientador: Prof. Dr. Alex Sandro Romeo de Souza Poletto
Área de Concentração: Bioinformática

ASSIS
2010

FICHA CATALOGRÁFICA

FRIZANCO, Renato A.
Um estudo sobre a bioinformática: Distinção dos tipos de dados e as diferentes bases de dados existentes / Renato Aparecido Frizanco. Fundação Educacional do Município de Assis - FEMA – Assis, 2010.
49p.

Orientador: Alex Sandro Romeo de Souza Poletto
Trabalho de Conclusão de Curso – Instituto Municipal de Ensino Superior de Assis – IMESA.

1.Bioinformática 2.Extração 3.Dados

CDD: 001.6
Biblioteca da FEMA

Fundação Educacional do Município de Assis
Instituto Municípal de Ensino Superior de Assis - IMESA

RENATO APARECIDO FRIZANCO

UM ESTUDO SOBRE A BIOINFORMÁTICA: DISTINÇÃO DOS TIPOS DE DADOS E AS DIFERENTES BASES DE DADOS EXISTENTES

Trabalho de Conclusão de Curso apresentado ao Instituto Municipal de Ensino Superior de Assis, como requisito do Curso de Graduação, analisado pela seguinte comissão examinadora:

Orientador: Prof. Dr. Alex Sandro Romeo de Souza Poletto

Analisador(1):

Assis
2010

DEDICATÓRIA

Dedico não somente este trabalho, como também a pessoa que sou hoje, aos meus pais Ana e Aparecido Frizanco, e aos amigos que durante esses últimos quatro anos da minha vida, puderam compartilhar momentos de tristeza e alegria, sempre mostrando que posso ser um grande sonhador.

AGRADECIMENTO

Agradeço a todos os professores do curso de Ciência da Computação pelos conhecimentos repassados nesses últimos anos, todo apoio e compreensão de cada um que foi necessário para conclusão desse trabalho.

Ao orientador Professor Doutor Alex Poletto por todo apoio prestado para a elaboração.

Aos amigos, que para citar haveria uma lista imensa, agradeço aos que confiaram em mim e me ajudaram nas horas que mais precisei.

Agradeço ao Grupo Cocal pela oportunidade dada a mim em 2003, oportunidade essa que fez com que eu chegasse onde estou hoje. A todos da empresa que mesmo não sabendo que estavam me ajudando acabaram por auxiliar nessa batalha.

A todos que de alguma forma contribuíram para a realização do trabalho.

“Sempre que notar que está do lado da maioria, faça uma pausa para refletir.”

Mark Twain

(1835-1910)

RESUMO

Este trabalho tem por objetivo apresentar uma introdução à Bioinformática levantando os conceitos mais utilizados na área, com principal foco no uso de bancos de dados quanto à forma de armazenamento e recuperação das informações e apresentação de exemplos de aplicações atualmente utilizadas nessa área.

Palavras-chaves: bioinformática, extração, dados.

ABSTRACT

This work aims at presenting an introduction to Bioinformatics lifting concepts used in the area, with main focus on the use of databases in the storage and retrieval of information and examples of applications currently used in this area.

Keywords: bioinformatics, extraction, data.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

2TRX	Identificador da tioredoxina de E coli.
DDBJ	Banco de Dados de DNA do Japão
DNA	Ácido DesoxirriboNucleico
EMBL	Laboratório Europeu de Biologia Molecular
PDB	*Protein Data Bank*
PGH	Projeto Genoma Humano
PIR	*Protein Information Resource*
RCSB	*Research Collaboratory for Structural Bioinformatics*
RNA	Ácido RiboNucleico
SGBD	Sistema Gerenciador de Banco de Dados
WWPDB	*Worldwide Protein Data Bank*
XML	*Extensible Markup Language*

SUMÁRIO

1. INTRODUÇÃO

Bioinformática é definida como um ponto comum entre a Biologia e a Computação, sendo a arte de utilizar recursos tecnológicos de informática para auxiliar nos estudos relacionados à Biologia. Com o surgimento dessa área, os dados obtidos são cada vez mais elevados, e a necessidade de armazenamento dessas informações tem sido um desafio para a computação (LIFSCHITZ, 2007).

Existem equipamentos muito sofisticados que auxiliam na extração de informações biológicas, porém mesmo que se tenham recursos muitos avançados para essas pesquisas, a forma na qual os cientistas ainda realizam suas buscas deixa a desejar, pelo fato da lentidão que é encontrado.

Os Bancos de Dados Relacionais atuais tem sido utilizados em maior escala para gerenciamento de dados empresariais, na qual é gravado nesses bancos tipos de dados simples, como números, caracteres ou datas. Poucos bancos possuem a funcionalidade de tratar dados complexos e variados como a biologia pode oferecer com os resultados de pesquisas (BANERJEE, 2000).

Para definir uma ferramenta para armazenamento desses dados, um ponto que deve ser analisado é a robustez na qual esses bancos de dados devem ser implementados, bem com a necessidade de um ótimo Sistema Gerenciador de Banco de Dados (SGBD).

Para demonstrar essa necessidade, tem-se como exemplo o Projeto Genoma Humano (PGH), projeto que teve início em 1990 e dado como concluído em 2003. Após o término desse projeto que visava seqüenciar o código genético humano, foi constatado que a sequência completa de DNA de um humano tem cerca de três bilhões de bases (OLIVEIRA, 2010; LIFSCHITZ, 2007).

Outro exemplo que pode ser citado é a necessidade do arquivamento mundial de sequencias de ácidos nucleicos, na qual se trata de uma parceria tríplice entre o National Center for Biotechnology Information (Estados Unidos), o EMBL Data Library (European Bioformatics Institute, Reino Unido) e o DNA Data Bank of Japan (National Institute of Genetics, Japão). Esses grupos trocam informações

diariamente. Embora o formato na qual esses dados seja armazenados e a natureza da anotação não serem idênticos, os dados brutos devem ser.

Esses bancos de dados organizam, arquivam e distribuem sequencias de DNA e RNA coletadas de projetos, publicações científicas e depósitos patentes.

Tomando com base todo esse panorama, serão realizados estudos de como esses dados estão sendo armazenados, incluindo nesse estudo, tipos de bases, exemplos de aplicações e tipos de dados existentes a fim de mapear as dificuldades encontradas pelos laboratórios biológicos.

1.1 OBJETIVOS

O objetivo principal desse trabalho é apresentar uma introdução à Bioinformática, levantando os conceitos mais utilizados na área em relação a banco de dados e as aplicações já existentes no mercado. Levantar os tipos de dados biológicos e a forma como os pesquisadores vêm armazenando, e os recursos utilizados para a recuperação dos mesmos para possibilitar melhores progressos das pesquisas biológicas.

1.2 JUSTIFICATIVAS

O alto volume de informações adquiridas pelos laboratórios de biologia e a dificuldade de realizar a comparação desses resultados é o que justifica a necessidade do levantamento desses conceitos e tipos de dados para adquirir o conhecimento dos distintos dados que precisam de tratamento no armazenamento e recuperação das informações.

1.3 MOTIVAÇÕES

A grande necessidade da padronização dos dados no mercado de trabalho é o que motiva a realização desse trabalho. A forma como o resultado irá impactar na área de bioinformática será notavelmente considerável.

1.4. ESTRUTURA DO TRABALHO

Este trabalho foi organizado em seis capítulos, sendo o primeiro esta introdução.

No segundo capítulo, serão apresentadas as fundamentações teóricas sobre Bioinformática.

No terceiro capítulo, serão apresentados tipos de bancos de dados biológicos existente.

No quarto capítulo, serão apresentados alguns exemplos de bancos de dados disponíveis mundialmente para estudo.

No quinto capítulo serão apresentadas aplicações que auxiliam nos estudos dos dados disponíveis nas bases que também serão citadas.

No sexto capítulo, serão apresentadas as considerações finais e as projeções futuras.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo será feita uma descrição de toda fundamentação teórica.

2.1. BIOINFORMÁTICA

Com o avanço das áreas de Biologia e Informática há um ponto em que ambas se encontram, surgindo a partir de então a Bioinformática.

Surge essa ciência devido à necessidade de ferramentas e equipamentos mais sofisticados de alta precisão para analisar o crescente volume de dados gerados em biologia molecular.

Ela tem como objetivo o gerenciamento e a análise de dados biológicos usando técnicas avançadas da computação a partir de desenvolvimento e implementação de ferramentas que possibilitem o gerenciamento e acesso eficientes de vários tipos de informações.

Nesse trabalho será mostrado o quanto essa ciência auxilia na exploração de dados genômicos.

2.2. GENOMA HUMANO

Pode-se dizer que Genoma Humano é o código genético do ser humano presente em cada uma das células, ou seja, o conjunto dos genes humanos. Nesse material são encontradas todas as informações para o funcionamento do organismo. É onde são gravadas nossas características hereditárias encarregadas de dirigir o desenvolvimento biológico de cada indivíduo.

Todas as informações são codificadas pelo DNA, o ácido desoxirribonucléico. Este, que tem um formato de dupla hélice é formado por quatro bases que se juntam aos pares, adenina com timina e citosina com guanina. Ver Figura 1 (RIDLEY, 2001).

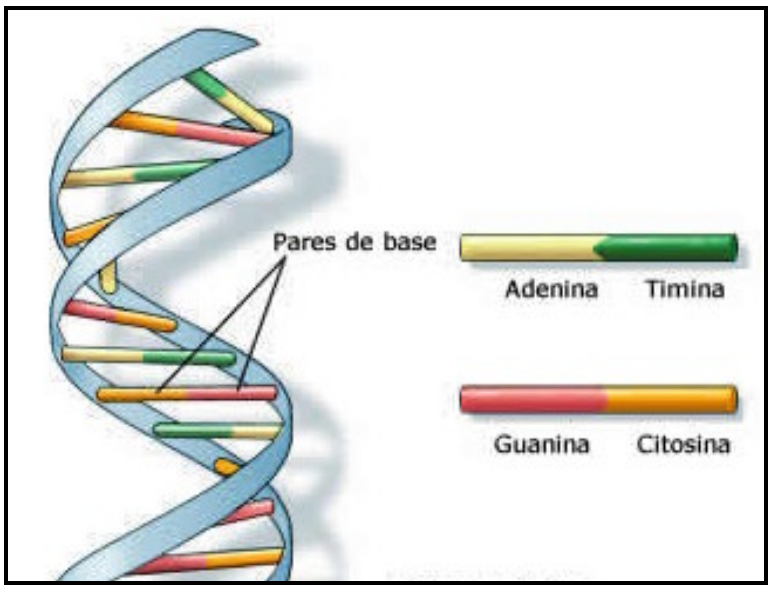


Figura 1. O DNA e suas ligações.

A importância do estudo das informações do genoma humano é que por intermédio desse mapeamento possamos descobrir a causa de muitas doenças antes mesmo delas surgirem e possamos adotar medidas de prevenções.

2.2.1. O PROJETO GENOMA HUMANO

O Projeto Genoma Humano (PGH) é um empreendimento internacional, que foi projetado para uma duração de quinze anos. Teve início em 1990 com vários objetivos, entre eles identificar e fazer o mapeamento dos cerca de 80 mil genes que se calculava existirem no DNA das células do corpo humano. Determinar as seqüências dos três bilhões de bases químicas que compõe o DNA humano. Armazenar essa informação em bancos de dados, desenvolverem ferramentas eficientes para analisar esses dados e torná-los acessíveis para novas pesquisas biológicas.

Outro objetivo do PGH foi descobrir todos os genes na seqüência de DNA e desenvolver meios de usar esta informação no estudo da Biologia e da Medicina, envolvendo com isso a melhoria e simplificação dos métodos de diagnósticos de doenças genéticas, otimização das terapêuticas para essas doenças e prevenção de doenças multifatoriais, no que diz respeito à saúde (RIDLEY, 2001).

2.3. ALINHAMENTO DE SEQUÊNCIAS

Na área de bioinformática, o alinhamento de sequências biológicas é um passo obrigatório para o avanço de pesquisas, passo inicial que trata a forma de realizar as análises de regiões conservadoras e de regiões que sofreram mutações em sequencias homólogas. Servindo também como ponto de partida para outras aplicações em biologia computacional, como o estudo de estruturas secundárias de proteínas.

Dadas duas sequencias biológicas TAGGTCA e TAGCTA, na imagem abaixo é possível entender o alinhamento de sequencias. (BRITO, 2003)

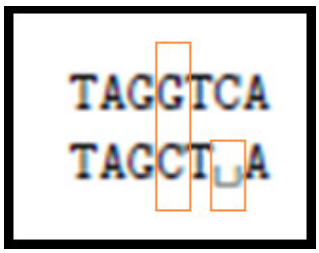


Figura 2. Resultado do alinhamento de sequências

Espaçamentos, também chamados de Gaps, podem ser inseridos entre as sequencias para que caracteres semelhantes sejam alinhados em colunas sucessivas.

Após essa inserção entre T e A na segunda sequência do exemplo citado podemos visualizar que a única diferença entre ele é a base G da primeira sequência que ocupa o lugar da base C da segunda.

Aplicações para essa tarefa devem existir para o auxilio dos pesquisadores.

2.4. ANOTAÇÃO GENÔMICA

Uma anotação trata-se de um dado extraído através da interpretação de um determinado experimento. Consiste na identificação das regiões codificantes das sequencias biológicas.

Para executar essa rotina é necessário que existam ferramentas que auxiliem nesse trabalho.

Uma anotação pode ser definida em três tipos, sendo elas:

- **Manual –** Na qual se trata de informações criadas pelo próprio pesquisador, que efetua pesquisas na literatura ou utiliza o próprio conhecimento para realizar as associações manualmente.
- **Automática –** Informação gerada por programas de análise ou importada de fontes de dados públicas. (BELLOZE, 2007)
- **Semi-automatizada –** Automáticas acrescidas de informações observadas pelos pesquisadores.

2.5. BIOLOGIA MOLECULAR

É o estudo da biologia em nível molecular, na qual consiste em estudar as interações entre os sistemas celulares, a partir da relação entre o DNA, RNA e a síntese das proteínas, e como essas interações são reguladas.

Assim, a essência da biologia molecular compreende o estudo dos processos de replicações, transcrição e tradução do material genético e a regulação desses processos. São frequentemente combinadas técnicas provindas da microbiologia, genética, bioquímica e biofísica. (LEHNINGER, 1995)

3. TIPOS DE BANCOS DE DADOS BIOLÓGICOS

A principal característica da genética médica atual é a crescente utilização da análise direta do material genético, tanto para diagnósticos quanto para pesquisas. A estocagem das amostras de DNA origina os Bancos de Dados. Podem-se diferenciar quatro tipos de Bancos de Material Genético, de acordo com suas características: de pesquisa, de diagnóstico, de dados e potenciais.

Os bancos de pesquisas são formados por dados obtidos de indivíduos ou de famílias extensas, e algumas vezes por populações inteiras, portadoras ou afetadas por uma determinada doença genética. Estes bancos podem ser organizados e mantidos por entidades públicas ou por empresas privadas.

Os bancos de diagnóstico são obtidos a partir de dados de pessoas com suspeita de determinada doença e de seus familiares, em geral para medidas de diagnósticos ou de aconselhamento (detecção de portadores, prognóstico).

Os bancos de dados de DNA são para que as informações sejam armazenadas para a identificação de um indivíduo por comparação com o padrão armazenado. Estes bancos geralmente têm caráter forense ou militar e várias críticas têm sido feitas quanto a sua utilização, tanto do ponto de vista tecnológico quanto ético, na qual um dos principais problemas diz respeito à privacidade e autonomia dos indivíduos analisados.

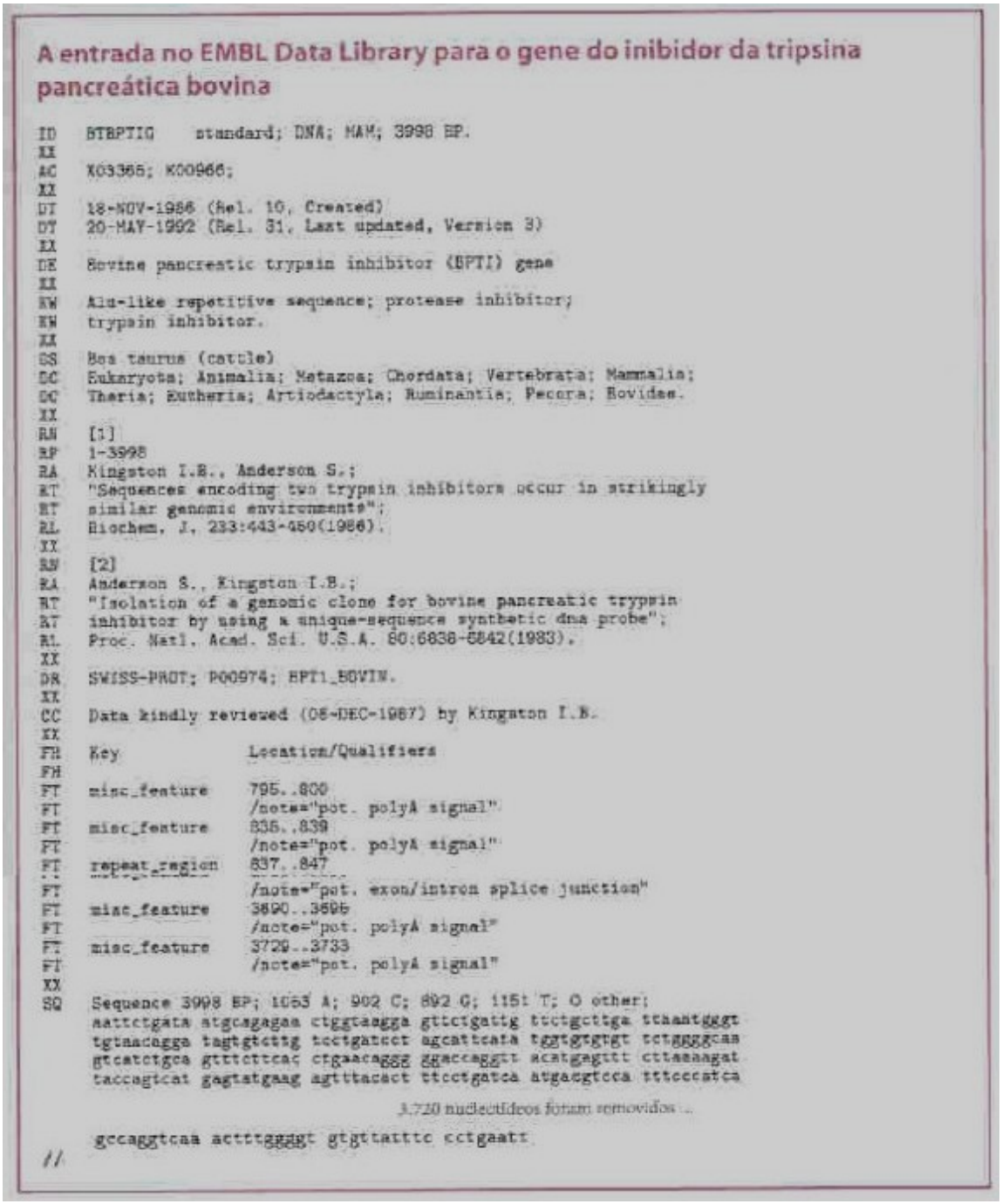
Os bancos potenciais de material genético são formados por qualquer coleção de tecido genético em geral.

Dos tipos citados acima, tem-se os Bancos de Dados que serão apresentados no próximo capítulo.

3.1. BANCO DE DADOS DE SEQUENCIAS DE ÁCIDOS NUCLÉICOS

Os bancos de dados de ácidos nucleicos, como distribuídos, são coleções de registros ou entradas. Cada entrada tem a forma de um arquivo de texto contendo dados e anotações para uma sequencia contígua única. Muitas entradas são reunidas a partir de diversos artigos publicados que descrevem fragmentos sobrepostos de uma sequencia completa. (LESK, 2008)

As entradas possuem um ciclo de vida no banco de dados. A necessidade dos usuários é poder ter acesso rápido aos dados armazenados, porém antes que as anotações estejam completas e que verificações sejam realizadas novas entradas podem ser disponibilizadas. As entradas passam pelas seguintes classes até chegarem ao formato final.



A entrada no EMBL Data Library para o gene do inibidor da tripsina pancreática bovina

```
ID   BTBPTIG    standard; DNA; MAM; 3998 BP.
XX
AC   X03366; K00966;
XX
DT   18-NOV-1986 (Rel. 10, Created)
DT   20-MAY-1992 (Rel. 31, Last updated, Version 3)
XX
DE   Bovine pancreatic trypsin inhibitor (BPTI) gene
XX
KW   Alu-like repetitive sequence; protease inhibitor;
KW   trypsin inhibitor.
XX
OS   Bos taurus (cattle)
OC   Eukaryota; Animalia; Metazoa; Chordata; Vertebrata; Mammalia;
OC   Theria; Eutheria; Artiodactyla; Ruminantia; Pecora; Bovidae.
XX
RN   [1]
RP   1-3998
RA   Kingston I.B., Anderson S.;
RT   "Sequences encoding two trypsin inhibitors occur in strikingly
RT   similar genomic environments";
RL   Biochem. J. 233:443-450(1986).
XX
RN   [2]
RA   Anderson S., Kingston I.B.;
RT   "Isolation of a genomic clone for bovine pancreatic trypsin
RT   inhibitor by using a unique-sequence synthetic dna probe";
RL   Proc. Natl. Acad. Sci. U.S.A. 80:6838-6842(1983).
XX
DR   SWISS-PROT; P00974; BPT1_BOVIN.
XX
CC   Data kindly reviewed (08-DEC-1987) by Kingston I.B.
XX
FH   Key             Location/Qualifiers
FH
FT   misc_feature    795..800
FT                   /note="pot. polyA signal"
FT   misc_feature    835..839
FT                   /note="pot. polyA signal"
FT   repeat_region   837..847
FT                   /note="pot. exon/intron splice junction"
FT   misc_feature    3690..3695
FT                   /note="pot. polyA signal"
FT   misc_feature    3729..3733
FT                   /note="pot. polyA signal"
XX
SQ   Sequence 3998 BP; 1065 A; 902 C; 892 G; 1151 T; 0 other;
     aattctgata atgcagagaa ctggtaagga gttctgattg ttctgcttga ttaaatgggt
     tgtaacagga tagtgtcttg tcctgatcct agcattcata tggtgtgtgt tctggggcaa
     gtcatctgca gtttcttcac ctgaacagag ggaccaggtt acatgagttt cttaaaagat
     taccagtcat gagtatgaag agtttacact ttcctgatca atgacgtcca tttcccatca

                                  3.720 nucleotídeos foram removidos ...

     gccaggtcaa actttgggat gtgttatttc cctgaatt
//
```

Figura 3. Entrada no EMBL Data Library.

Não anotada ➔ Preliminar ➔ Não revisada ➔ Padrão

Raramente uma entrada é removida quando se determinou que estivessem incorretas.

Uma amostra de uma entrada de sequencia de DNA do EMBL Data Library, incluindo anotações assim como os dados da sequencia, é o gene inibidor da tripsina pancreática bovina.

A Figura 3 mostra parte dessa entrada, omitindo a maior parte da sequencia propriamente dita. As linhas que se iniciam por FT, de Feature Tables, seriam as tabelas de características na qual define propriedades de regiões especificas, por exemplo, sequencias codificadoras. Essas sequências são planejadas para serem lidas por aplicações específicas. Elas têm um formato controlado e um vocabulário mais restrito. (LESK, 2008)

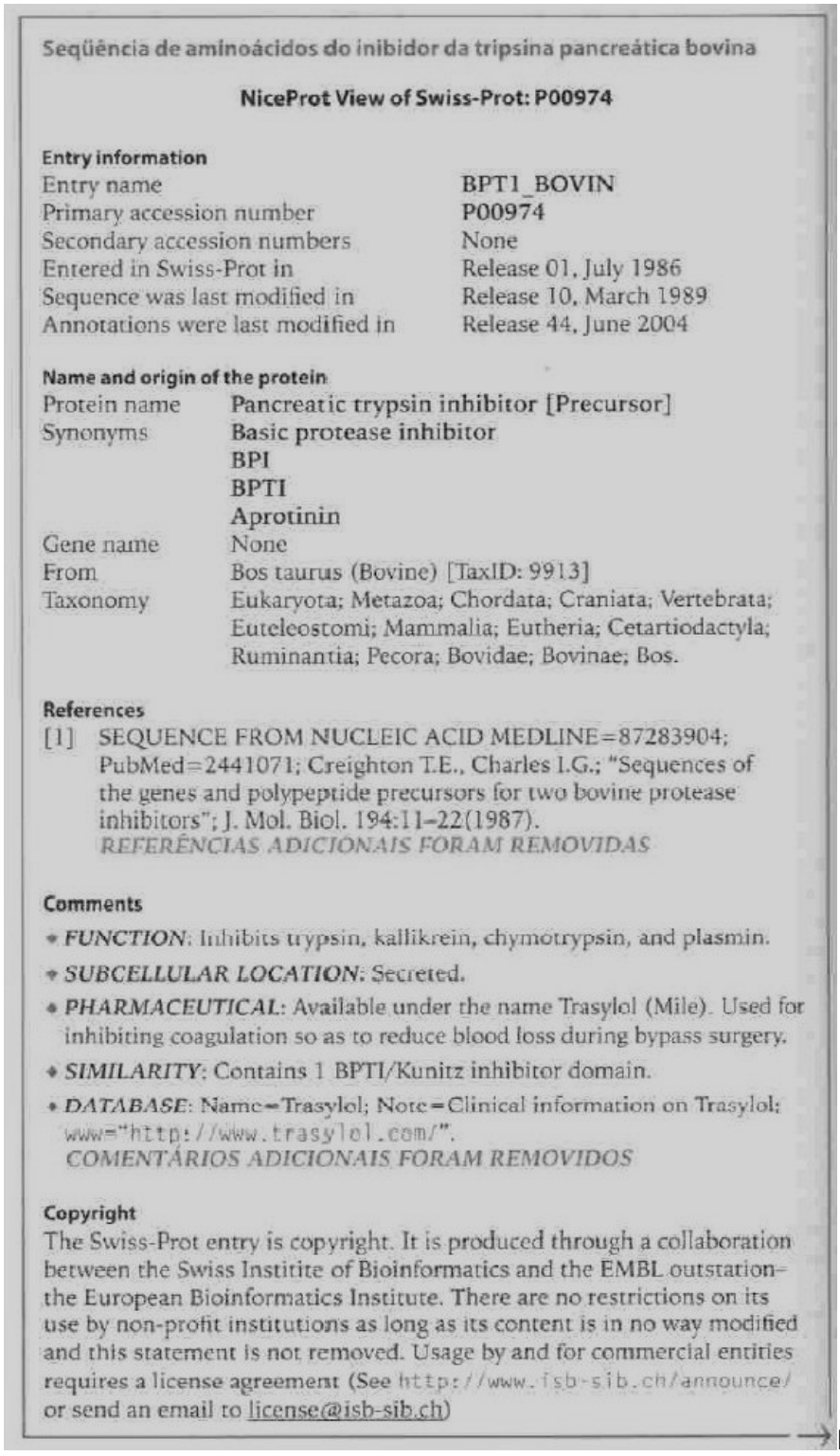
Na seção 4.3 será mostrado como essas entradas são inseridas na base de dados.

O desenvolvimento de vocabulários controlados e de um dicionário-padrão e sinônimos compartilhados, para palavra-chave e tabelas de características, são importantes também para estabelecer conexões entre os diferentes bancos de dados.

3.2. BANCO DE DADOS DE SEQÜÊNCIAS DE PROTEÍNAS

Em 2002, três bancos de dados de proteínas, o *Protein Information Resource* - PIR no *National Biomedical* em *Washigton*, DC, Estados Unidos, o SWISS-PROT e TREMBL, do *Swiss Institute of Bioinformatcs*, Genebra, e o *European Bioinformatics Institute* em *Hinxton*, Reino Unido, coordenaram seu esforços para formar uma parceria chamada de Consorcio UnitProt. Os parceiros dessa iniciativa compartilharam o banco de dados, mas continuam a oferecer ferramentas separadas de acesso à recuperação de informação. (LESK, 2008)

Proveniente do UnitProt, a entrada para a sequencia de aminoácidos do gene inibidor da tripsina pancreática bovina, no formato SWISS-PROT, é mostrado nas imagens a seguir.



Seqüência de aminoácidos do inibidor da tripsina pancreática bovina

NiceProt View of Swiss-Prot: P00974

Entry information

Entry name	BPT1_BOVIN
Primary accession number	P00974
Secondary accession numbers	None
Entered in Swiss-Prot in	Release 01, July 1986
Sequence was last modified in	Release 10, March 1989
Annotations were last modified in	Release 44, June 2004

Name and origin of the protein

Protein name	Pancreatic trypsin inhibitor [Precursor]
Synonyms	Basic protease inhibitor BPI BPTI Aprotinin
Gene name	None
From	Bos taurus (Bovine) [TaxID: 9913]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Cetartiodactyla; Ruminantia; Pecora; Bovidae; Bovinae; Bos.

References

[1] SEQUENCE FROM NUCLEIC ACID MEDLINE=87283904; PubMed=2441071; Creighton T.E., Charles I.G.; "Sequences of the genes and polypeptide precursors for two bovine protease inhibitors"; J. Mol. Biol. 194:11–22(1987).
REFERÊNCIAS ADICIONAIS FORAM REMOVIDAS

Comments

- ***FUNCTION***: Inhibits trypsin, kallikrein, chymotrypsin, and plasmin.
- ***SUBCELLULAR LOCATION***: Secreted.
- ***PHARMACEUTICAL***: Available under the name Trasylol (Mile). Used for inhibiting coagulation so as to reduce blood loss during bypass surgery.
- ***SIMILARITY***: Contains 1 BPTI/Kunitz inhibitor domain.
- ***DATABASE***: Name=Trasylol; Note=Clinical information on Trasylol; www="http://www.trasylol.com/".
COMENTÁRIOS ADICIONAIS FORAM REMOVIDOS

Copyright

The Swiss-Prot entry is copyright. It is produced through a collaboration between the Swiss Institite of Bioinformatics and the EMBL outstation–the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See http://www.isb-sib.ch/announce/ or send an email to license@isb-sib.ch)

Figura 4. Parte 1 da seqüência de aminoácidos no formato SWISS-PROT.

Cross-references

EMBL	M20934; AAD13685.1;-.
	REFERÊNCIAS CRUZADAS ADICIONAIS AO EMBL FORAM REMOVIDAS
PIR	S00277; TIBO.
PDB	1K09; 10-JUL-02.
	REFERÊNCIAS CRUZADAS ADICIONAIS AO PDB FORAM REMOVIDAS
InterPro	IPR002223; Kunitz_BPTI.
Pfam	PF00014; Kunitz_BPTI; 1.
	Pfam graphical view of domain structure.
PRINTS	PR00759; BASICPTASE.
ProDom	PD000222; Kunitz_BPTI; 1.
	[Domain structure/List of seq. sharing at least 1 domain]
SMART	SM00131; KU; 1.
PROSITE	PS00280; BPTI_KUNITZ_1; 1.
	PS50279; BPTI_KUNITZ_1; 2.
	PROSITE graphical view of domain structure.
Implicit links to	HOVERGEN; BLOCKS; ProtoNet; ProtoMap; PRESAGE; DIP; ModBase; SMR; SWISS-2DPAGE; UniRef.

Keywords
Serine protease inhibitor; Signal; Pharmaceutical; 3D-structure.

Features

Key	From	To	Length	Description
SIGNAL	1	21	21	Potential.
PROPEP	22	35	14	
CHAIN	36	93	58	Pancreatic trypsin inhibitor.
PROPEP	94	100	7	
DOMAIN	40	90	51	BPTI/Kunitz inhibitor.
SITE	50	51	2	Reactive bond for trypsin.
DISULFID	40	90		
DISULFID	49	73		
DISULFID	65	86		
HELIX	38	41	4	
STRAND	53	59	7	
TURN	60	63	4	
STRAND	64	70	7	
STRAND	80	80	1	
HELIX	83	90	8	

Sequence information

Length: **100 AA** [This is the length of the unprocessed precursor]
Molecular weight: **10903 Da** [This is the MW of the unprocessed precursor]
CRC64: **6A778A4AD763FB19** [This is a checksum on the sequence]

```
        10         20         30         40         50         60
        |          |          |          |          |          |
MKMSRLCLSV ALLVLLGTLA ASTPGCDTSN QAKAQRPDFC LEPPYTGPCK ARIIRYFYNA
        70         80         90        100
        |          |          |          |
KAGLCQTFVY GGCRAKRNNF KSAEDCMRTC GGAIGPWENL
```

Figura 5. Parte 2 da seqüência de no formato SWISS-PROT.

3.3. BANCO DE DADOS DE ESTRUTURAS

Bancos de dados de estruturas arquivam, anotam e distribuem conjuntos de coordenadas atômicas. O principal banco de dados de estrutura de macromoléculas biológicas é o *Protein Data Bank* (PBD). (LESK, 2008)

Estes dados, geralmente obtidos por Cristalografia de raios X ou Ressonâncias Magnéticas Nucleares por biólogos e bioquímicos de todo o mundo. Há uma vasta quantidade de dados em 3-D disponível livremente para estudos.

Para manter a organização desses dados PDB, o *Research Collaboratory for Structural Bioinformatics* (RCSB), o *Molecular Structure Database* e *o European Bioinformatics Institute* e o *Protein Data Bank of Japan f*undaram *o Worldwide Protein Data Bank* (wwPDB). O objetivo do wwPDB é disponibilizar e garantir uma biblioteca unificada desses arquivo PDB. Quando se fundou o wwPDB havia apenas 7 estruturas de proteínas. Hoje esse número já se ultrapassa 12.000 arquivos de estruturas de proteínas e de ácidos nucleicos, dentre outras biomoléculas. Esses arquivos são encontrados no endereço web http://www.rcsb.org. (LESK, 2008)

A seguir será mostrada uma parte de uma estrutura do Protein Data Bank para uma estrutura denominada tioredoxina de E coli. Identificada no wwPDB pelo identificador 2TRX com a descrição “*Crystal structure of thioredoxin from Escherichia coli at 1.68 a Resolution*”. Esse identificador de quatro caracteres é atribuído a cada uma das estruturas depositadas. A única regra para a criação desses é que o primeiro caractere seja um numero de 1 a 9, os outros não representam necessariamente siglas das fórmulas que compõem a estrutura. Em muitas situações diversos identificadores representam a mesma proteína, porém em estados de ligação distinto.

Na imagem seguinte é mostrado como uma Informação Sumarizada é mostrada no portal.

Summary | Sequence | Derived Data | Seq. Similarity | 3D Similarity | Literature | Biol. & Chem. | Methods | Geometry | Links

CRYSTAL STRUCTURE OF THIOREDOXIN FROM ESCHERICHIA COLI AT 1.68 ANGSTROMS RESOLUTION

DOI:10.2210/pdb2trx/pdb

Primary Citation

Crystal structure of thioredoxin from Escherichia coli at 1.68 A resolution.

Katti, S.K., LeMaster, D.M., Eklund, H.

Journal: (1990) J.Mol.Biol. **212:** 167-184

PubMed: 2181145
Search Related Articles in PubMed

PubMed Abstract:

The crystal structure of thioredoxin from Escherichia coli has been refined by the stereochemically restrained least-squares procedure to a crystallographic R-factor of 0.165 at 1.68 A resolution. In the final model, the root-mean-square deviation from ideality for bond distances is... [Read More & Search PubMed Abstracts]

Molecular Description Hide

Classification: Electron Transport
Structure Weight: 24329.32

Molecule: THIOREDOXIN
Polymer: 1 **Type:** polypeptide(L) **Length:** 108
Chains: A, B

Source Hide

Polymer: 1

Scientific Name: Escherichia coli Taxonomy

Ligand Chemical Component Hide

Identifier	Name	Formula	Binding Affinity (BindingDB)	Interaction View
CU	COPPER (II) ION	Cu		Ligand Explorer
MPD	(4S)-2-METHYL-2,4-PENTANEDIOL	$C_6 H_{14} O_2$		Ligand Explorer

Derived Data Hide

- SCOP Classification v1.75: (2 Domains) - (SCOP)
- CATH Classification v3.3.0: (2 Domains) - (CATH)
- PFAM Classification: 2 Domains - (PFAM)
- GO Terms: 3 Terms - (GO)

Figura 6. Página sumária para a entrada 2TRX, a da tioredoxina de E.Coli.

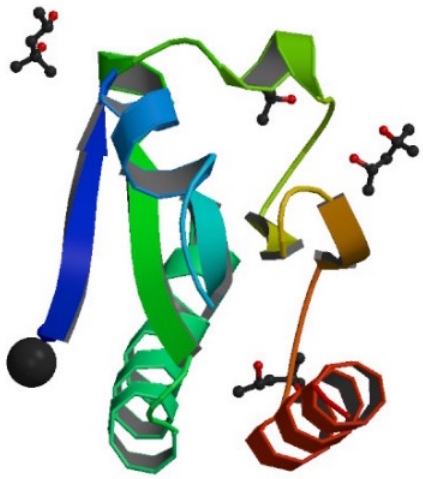


Figura 7. Montagem da imagem biológica da entrada 2TRX.

A partir desse identificador na base obtemos todas as entradas de átomos de proteína, e milhares cadeias de moléculas. No exemplo abaixo foram removidas várias dessas entradas.

```
HEADER    ELECTRON TRANSPORT                      19-MAR-90   2TRX
TITLE     CRYSTAL STRUCTURE OF THIOREDOXIN FROM ESCHERICHIA COLI AT
TITLE    2 1.68 ANGSTROMS RESOLUTION
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: THIOREDOXIN;
COMPND   3 CHAIN: A, B;
COMPND   4 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: ESCHERICHIA COLI;
SOURCE   3 ORGANISM_TAXID: 562
JRNL        TITL   CRYSTAL STRUCTURE OF THIOREDOXIN FROM ESCHERICHIA
JRNL        TITL 2 COLI AT 1.68 A RESOLUTION.
JRNL        REF    J.MOL.BIOL.                   V. 212   167 1990
JRNL        REFN                   ISSN 0022-2836
JRNL        PMID   2181145
JRNL        DOI    10.1016/0022-2836(90)90313-B
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH   A.HOLMGREN,B.-O.SODERBERG,H.EKLUND,C.-I.BRANDEN
REMARK   1  TITL   THREE-DIMENSIONAL STRUCTURE OF ESCHERICHIA COLI
REMARK   1  TITL 2 THIOREDOXIN-S2 TO 2.8 ANGSTROMS RESOLUTION
REMARK   1  REF    PROC.NATL.ACAD.SCI.USA        V.  72  2305 1975
REMARK   1  REFN                   ISSN 0027-8424
REMARK   1 REFERENCE 2
REMARK   1  AUTH   B.-O.SODERBERG,A.HOLMGREN,C.-I.BRANDEN
REMARK 200  DETECTOR MANUFACTURER          : NULL
REMARK 200  INTENSITY-INTEGRATION SOFTWARE : NULL
REMARK 200  DATA SCALING SOFTWARE          : NULL
SEQRES   3 B  108  PHE TRP ALA GLU TRP CYS GLY PRO CYS LYS MET ILE ALA
SEQRES   4 B  108  PRO ILE LEU ASP GLU ILE ALA ASP GLU TYR GLN GLY LYS
SEQRES   5 B  108  LEU THR VAL ALA LYS LEU ASN ILE ASP GLN ASN PRO GLY
SHEET    2 B1B 5 LEU B  53  ASN B  59  1  O  VAL B  55   N  ILE B   5
SHEET    3 B1B 5 GLY B  21  TRP B  28  1  N  TRP B  28   O  LEU B  58
SHEET    4 B1B 5 PRO B  76  LYS B  82 -1  O  THR B  77   N  PHE B  27
SHEET    5 B1B 5 VAL B  86  GLY B  92 -1  O  ALA B  87   N  LEU B  80
SSBOND   1 CYS A   32    CYS A   35                          1555   1555  2.09
CISPEP   2 ILE B   75    PRO B   76          0        -2.42
SITE     1 AC1  5 SER A   1  ASP A   2  LYS A   3  ASP A  10
SITE     2 AC1  5 HOH A 405
SITE     1 AC2  5 SER B   1  ASP B   2  LYS B   3  ASP B  10
SITE     2 AC2  5 HOH B 478
SITE     1 AC3  4 ASP A  10  ASP A  43  GLU A  44  HOH A 442
SITE     1 AC4  6 GLU A  44  HOH A 524  GLU B  30  TRP B  31
SITE     2 AC4  6 GLY B  33  LYS B  36
SITE     1 AC5  5 TYR B  70  ILE B  72  THR B  77  THR B  89
SITE     2 AC5  5 VAL B  91
SITE     1 AC6  3 ILE B  60  ALA B  67  ILE B  72
SITE     1 AC7  4 MET A  37  ILE A  38  ALA A  93  LEU A  94
SITE     1 AC8  4 TYR A  70  GLY A  71  THR A  89  VAL A  91
SITE     1 AC9  8 ILE A  60  ALA A  67  ILE A  72  ARG A  73
SITE     2 AC9  8 GLY A  74  ILE A  75  HOH A 494  HOH B 528
CRYST1   89.500   51.060   60.450  90.00 113.50  90.00 C 1 2 1        8
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.011173  0.000000  0.004858        0.00000
SCALE2      0.000000  0.019585  0.000000        0.00000
```

Figura 8. Entradas de átomos de proteína e cadeias de moléculas.

Nessa mesma página é disponibilizado um arquivo para download com a extensão PDB podendo ser aberto por aplicações específicas para uma melhor análise. É possível visualizar o ciclo de cada componente e explorar as subsequências por meio dessas aplicações, no capítulo a seguir será demonstrado um exemplo de uma aplicação com essa funcionalidade.

3.4. BANCO DE DADOS DE VIAS METABÓLICAS

A *Kyoto Encyclopedia of Genes and Genomes* (KEGG) arquiva genomas individuais, produtos de genes e suas funções, mas o diferencial principal está na sua integração de informações bioquímicas e genéticas. A KEGG se concentra nas interações associações de moléculas e redes metabólicas e reguladoras. Ela esta sendo desenvolvida sob a direção de M. Kanehisa. (LESK, 2008)

A KEGG organiza cinco tipos de dados em um sistema que compreende:

- Catálogos de compostos químicos em células vivas;
- Catálogos de genes;
- Mapas de genomas;
- Mapas de vias;
- Tabelas de ortólogos;

Os catálogos de compostos químicos e genes contêm informações sobre moléculas ou sequencias especificas.

Mapas de genomas integram os próprios genes de acordo com as suas localizações nos cromossomos.

Mapas de vias descrevem redes em potencial de atividades moleculares, tanto metabólicas quanto reguladoras. Uma via metabólica é uma idealização correspondendo a um grande numero de possíveis cascatas metabólicas. Ela pode gerar uma via metabólica real de um organismo particular, alinhando as proteínas daquele organismo com enzimas dentro das vias de referencias.

Tabelas de ortólogos, genes que codificam proteínas correspondentes em diferentes organismos, liga a enzima a outras relacionadas, presentes em outros organismos. Isso permite a análise das relações entre as vias metabólicas de diferentes organismos.

4. EXEMPLOS DE BANCO DE DADOS

Será apresentado nesse capítulo exemplos de banco de dados biológicos existentes.

4.1. BANCO DE DADOS GENBANK

O GENBANK, criado no centro norte americano *National Institute of Health (NIH),* para informação biotecnológica foi um dos primeiros e ainda mais popular banco de dados para o depósito de sequências de DNA. É lá que pesquisadores de todo o mundo depositam as sequências de A's (Adeninas), C's (Citosinas), G's (Guaninas) e T's (Timinas) da molécula de DNA que obtêm ao seqüenciar o genoma dos mais diversos organismos.

Possui parceria com os laboratórios *DataBank of Japan (DDBJ)* e com o *European Molecular Biology Laboratory (EMBL), na qual realizam compartilhamento* das milhares de sequencias armazenadas.

No final dos anos 90 observou-se um crescimento exponencial do número de seqüências de biomoléculas depositadas no GenBank (Figura 9). Esse aumento teve início a partir de 1990, quando surgiram os seqüenciadores de DNA a laser, totalmente automatizados.

Em fevereiro de 2004 essas bases de dados já chegavam a aproximadamente 38 bilhões de bases em 32,5 milhões de registros de sequencias. Em 2009, um número de 86 bilhões nas 83 milhões de sequencias depositadas. (SOARES, 2010)

Na Figura 9 é mostrado o crescimento dos números de sequencias mostrando o grande aumento no final dos anos 90.

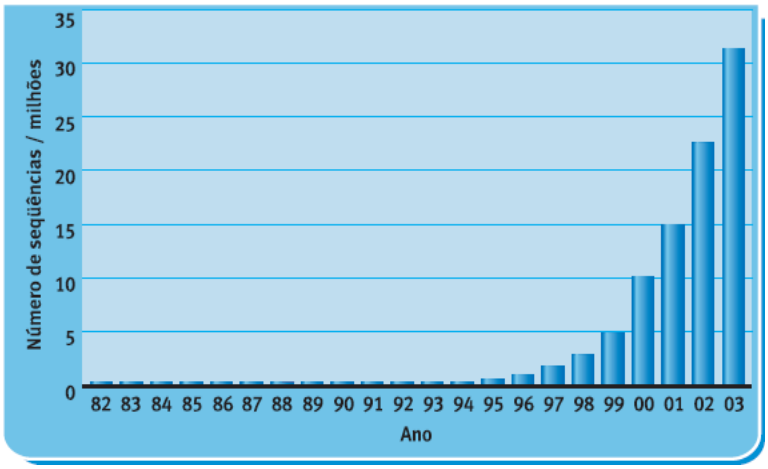


Figura 9. Numero de seqüências por milhões de biomoléculas.

Novas sequências são inseridas no GenBank a cada dois meses. Informações essas recebidas de laboratórios independentes e centros de sequenciamento em larga-escala do mundo inteiro. Esses dados são submetidos por e-mail destinados a gb-*admin@ncbi.nlm.nih.gov*, na qual devem ser enviados no formato de texto. Há um registro muito alto de arquivos que são devolvidos por apresentarem formato incorreto.

Uma das dificuldades encontradas é quanto à redundância de informações. Se uma entrada da característica estiver diferente com o que está armazenado na base e direcionado para inclusão de uma nova sequência não há uma forma de validar e essa é inserida.

Para não terem os arquivos rejeitados é importante que siga as características abaixo e que todas as entradas estejam consistentes:

- Descrição da Sequência
- Nome Científico e Taxonomia (Ciência de Classificar Organismos Vivos)
- Tabela de Característica
- Referencias Bibliográficas

Na imagem abaixo é mostrado como uma entrada deve ser enviada para atualização do GenBank. Milhares de linhas do atributo ORIGIN foram excluídos para resumo.

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"
                     /translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
                     TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
                     VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
                     VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE
                     TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTGNVSYDLPLNYV
                     YLDDDPISSDKLGSINLLDAPDWVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG
                     DVIYFNFEVVSTTDLFAISSLPNINATRGEWFSYYFLPSQFTDYVNTNVSLEFTNSSQ
                     DHDWVKFQSSNLTLAGEVPKNFDKLSLGLKANQGSQSQELYFNIIGMDSKITHSNHSA
                     NATSTRSSHHSTSTSSYTSSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA
                     CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDLNNPANKPNQENATPLN
                     NPFDDDASSYDDTSIARRLAALNTLKLDNHSATESDISSVDEKRDSLSGMNTYNDQFQ
                     SQSKEELLAKPPVQPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTGNLSPVSDIVRDS
                     YGSQKTVDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK
                     HRNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
                     VDFSNKSNVNVGQVKDIHGRIPEML"
     gene            complement(3300..4037)
                     /gene="REV7"
     CDS             complement(3300..4037)
                     /gene="REV7"
                     /codon_start=1
                     /product="Rev7p"
                     /protein_id="AAA98667.1"
                     /db_xref="GI:1293616"
                     /translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
                     FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
                     KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
                     RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
                     LISGDDKILNGVYSQYEEGESIFGSLF"
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
      481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
      541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
      601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt ccaacatcta
      661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacag
      721 ctactatatc actactccat ctagtagtgg ccacgcccta tgaggcatat cctatcggaa
```

Figura 10. Arquivo para atualização do GenBank (Saccharomyces cerevisiae)

4.2. BANCO DE DADOS DDBJ

O Banco de Dados de DNA do Japão, sob responsabilidade *do Center for Information Biology, National Institute of Genetics, Japan*, é um banco de dados de sequência de nucleotídeos certificado oficialmente para coletar sequências de nucleotídeos de pesquisadores e para o fornecimento desses dados para outros parceiros.

Assim como citado no capitulo anterior o DDBJ possui compartilhamento de informações com o *National Institute of Health*, responsável pelas informações contidas no GenBank.

O DDBJ vem desde 1988 recolhendo dados de sequência principalmente de pesquisadores japoneses, mas também aceita dados de pesquisadores de outros países.

A atualização da base é realizada por meio de uma ferramenta para envio dos arquivos disponível no próprio site do DDBJ. O formato das informações a serem enviadas é similar ao do GenBank. A validação é realizada no momento do upload da informação, mas não descarta a possibilidade de realizar inclusões de dados em redundância, assim como problema encontrado no GenBank.

Na imagem abaixo é exibido formulário a ser preenchido no momento do upload para atualização ou inclusão de novas sequencias no DDBJ.

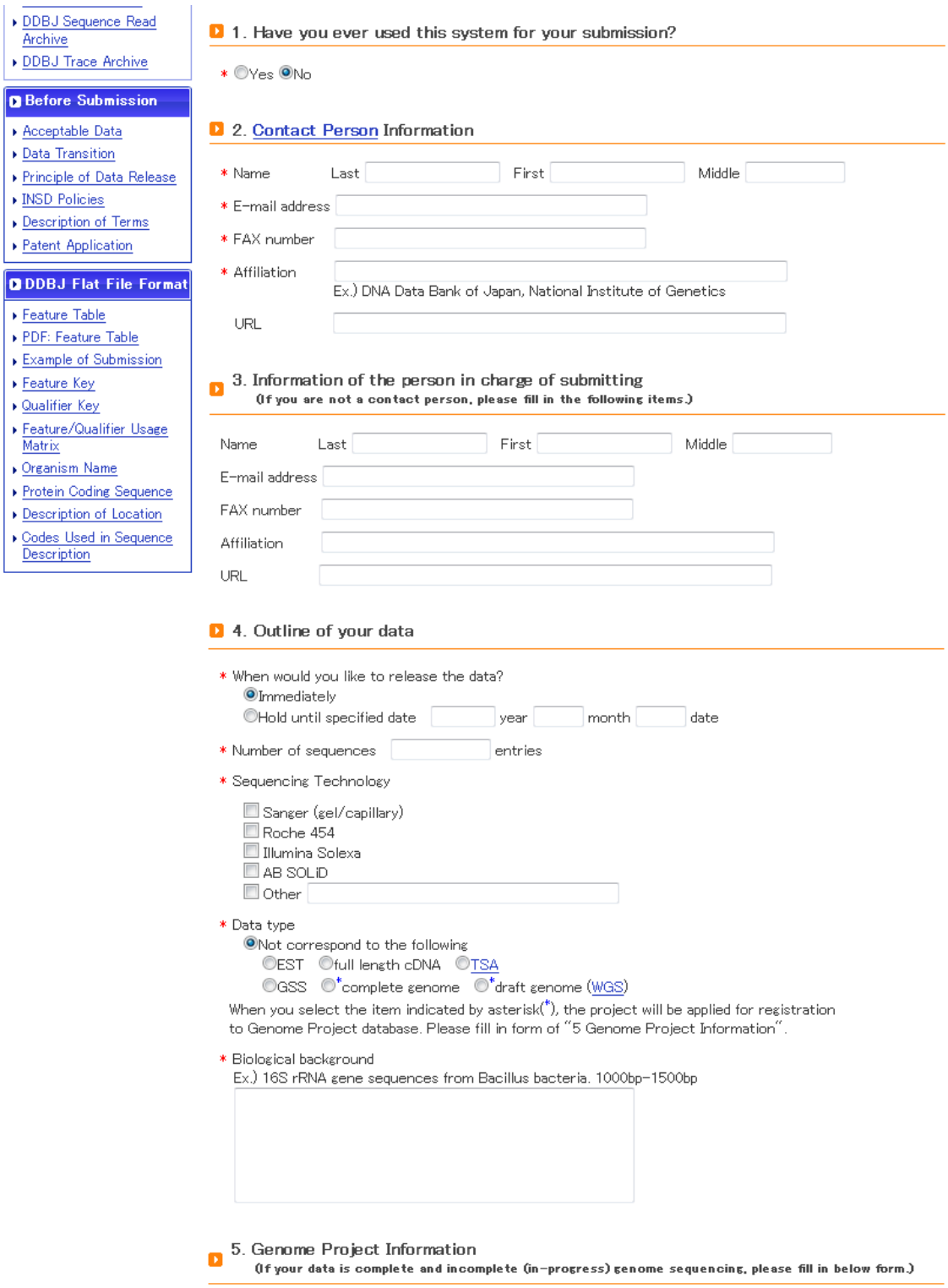


Figura 11. Formulário para submissão de dados no DDBJ

4.3. BANCO DE DADOS EMBL

O EMBL é um grande banco de dados responsável pelo armazenamento de sequencias primarias de nucleotídeos. Outra grande fonte de sequências de DNA e RNA. Desenvolvido por uma organização europeia liderada por Guy Cochrane e possui colaboração de outras organizações como o GenBank (EUA) e o banco de dados de DNA do Japão (DDBJ), compartilhando as informações. Cada um dos três grupos recolhe uma parcela dos dados da sequência do total registrado em todo o mundo, e todos os dados devem ser atualizados na base. Novas entradas devem ser trocadas entre os grupos em uma base diária.

Essa atualização é realizada por meio do site do EMBL similar ao processo para atualização do DDBJ, na qual necessita do preenchimento de um formulário para submissão, porem para acesso a esse formulário é necessário ser um laboratório anteriormente autorizado por eles. Aceitam também o recebimento dos arquivos de textos por e-mail, desde que os remetentes sejam os autorizados.

5. EXEMPLO DE APLICAÇÕES

Será apresentado nesse capitulo exemplos de sistemas existentes para o controle de informações biológicas. Os aplicativos aqui apresentados são todos desenvolvidos em Java sendo possível executar em qualquer plataforma que se tenha uma máquina virtual Java instalada. Java é uma linguagem de programação orientada a objetos, desenvolvida por uma equipe na *Sun Microsystens.* Uma linguagem que vem participando com grande atuação no mercado de trabalho.

5.1. JMOL– VISUALIZADOR 3D DE MOLÉCULAS

JMol é desenvolvido em Java, *open source*, gratuito, Destinado a estudantes, educadores e principalmente a pesquisadores de bioquímica.

O aplicativo permite visualizar estruturas moleculares simples ou complexas, como o DNA, de modo tridimensional. Capaz de demonstrar todas as características de quaisquer componentes que estejam disponíveis nos 49 tipos de arquivos possíveis que a aplicação executa. A mais conhecida das extensões válidas é a PDB, como descrito na seção 3.3., podendo ser encontrado no wwPDB.

Abaixo serão demonstradas algumas visualizações da composição tioredoxina de E coli identificada pelo identificador 2TRX

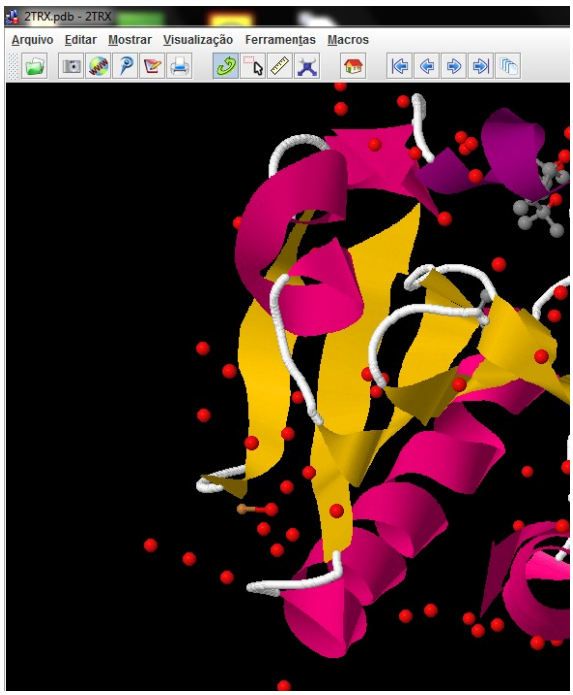


Figura 12. Imagem da composição da entrada 2TRX.

Na próxima imagem é exibido a composição de forma ampliada pelo aplicação

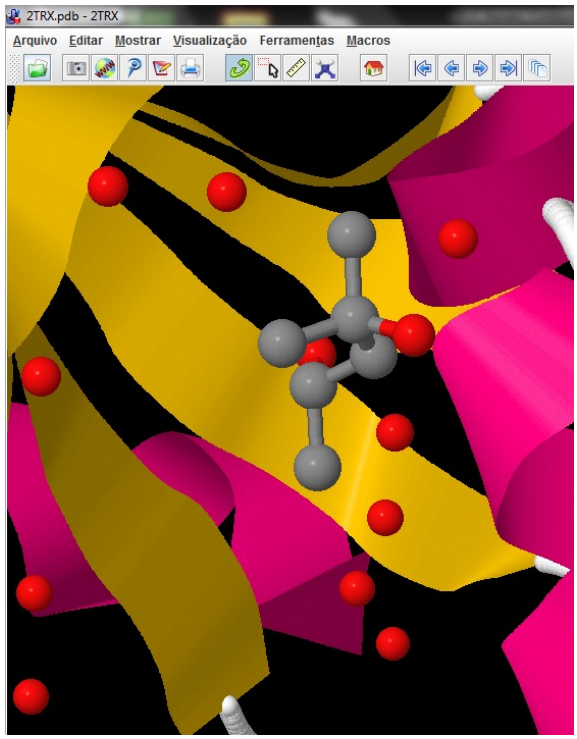


Figura 13. Imagem ampliada da composição da entrada 2TRX.

Na imagem acima cada cor representa um elemento da composição. Essas cores são definidas de acordo com padrões químicos e biológicos já definidos por cada ciência.

Quando se tratar de elementos químicos, por exemplo, as cores da imagem abaixo seguindo o padrão CPK (*CPK coloring*, sigla derivada dos *químicos Corey, R., Pauling, L. e Kolton, W.)* é o que foi definido no desenvolvimento da aplicação. Pelas cores abaixo cada cientista saberá de qual elemento a imagem se trata sem que precise exibir os rótulos, o que também é possível para iniciantes que não são acostumados com as cores pré-definidas.

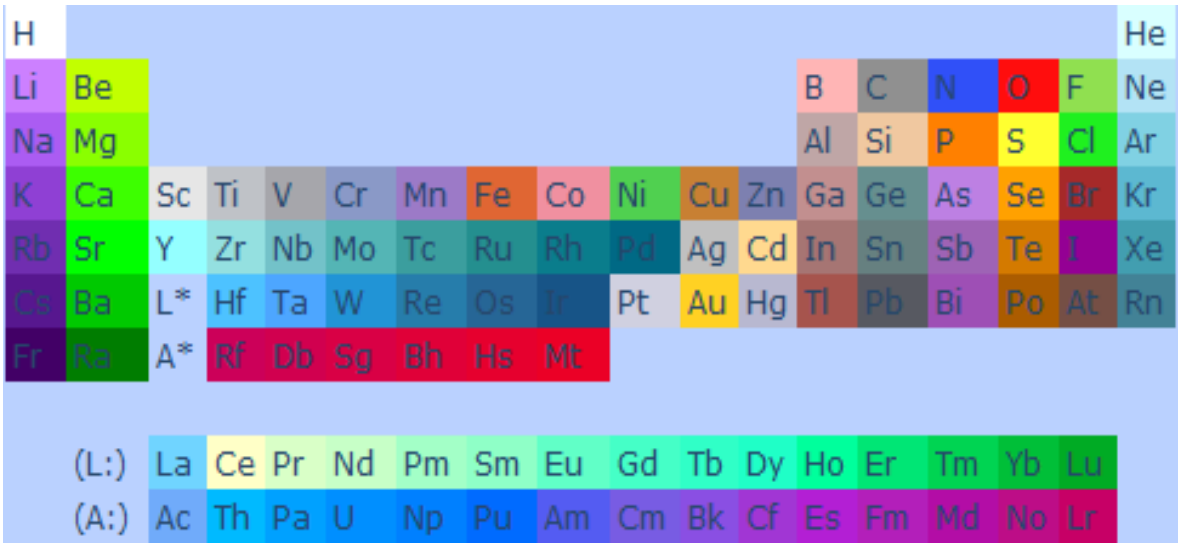


Figura 14. Código de cores dos elementos químicos (CPK)

A seguir serão mostrados outros padrões definidos, para as cargas elétricas, sequencias de aminoácidos e estruturas secundárias.

Cores por carga elétrica:

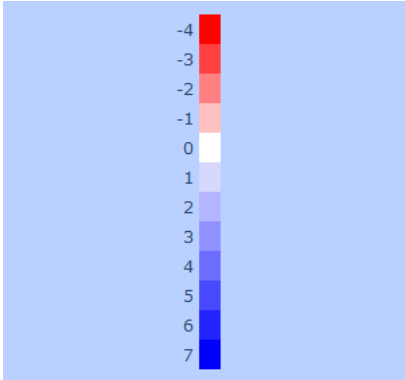


Figura 15. Cores por carga elétrica

Cores por aminoácidos ou nucleotídeos:

Figura 16. Cores por aminoácidos ou nucleotídeos

Cores por estrutura secundária:

Figura 17. Cores por estrutura secundária

As setas que são exibidas nas imagens referem-se às direções dos movimentos que as moléculas se comportam.

A seguir será exibida uma sequencia de imagem para uma determinada composição, onde na primeira imagem ilustra a posição inicial e na imagem seguinte a posição final.

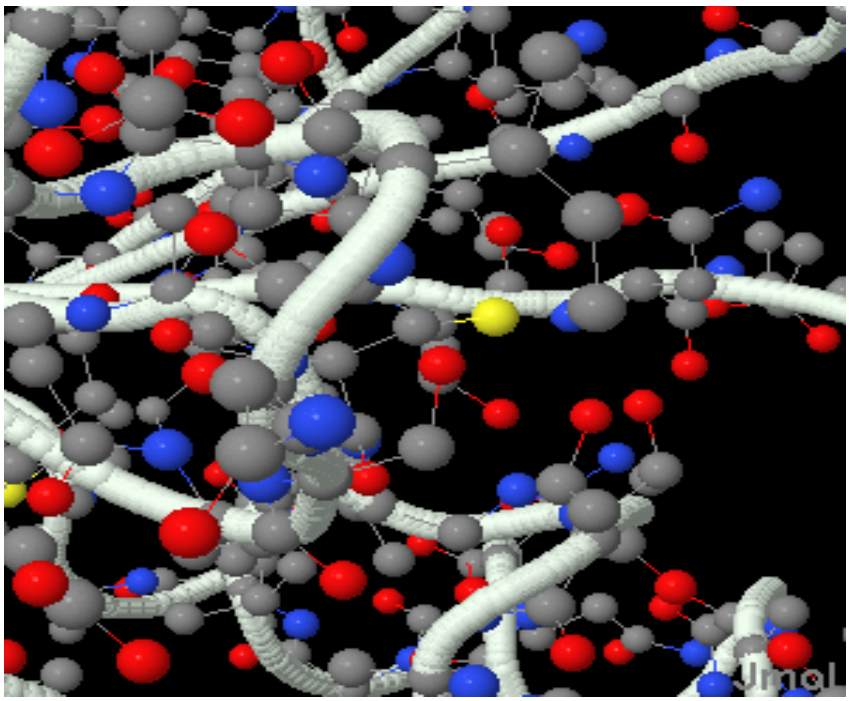


Figura 18. Posição inicial antes do movimento das moléculas

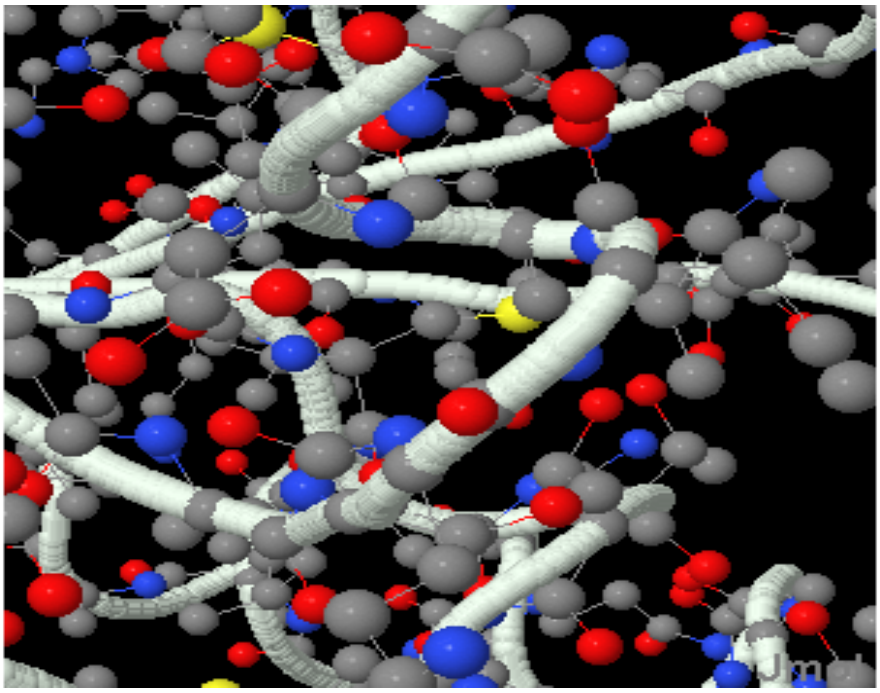


Figura 19. Posição final depois do movimento das moléculas

A imagem abaixo representa todos os movimentos realizados, identificados pelas setas citadas anteriormente.

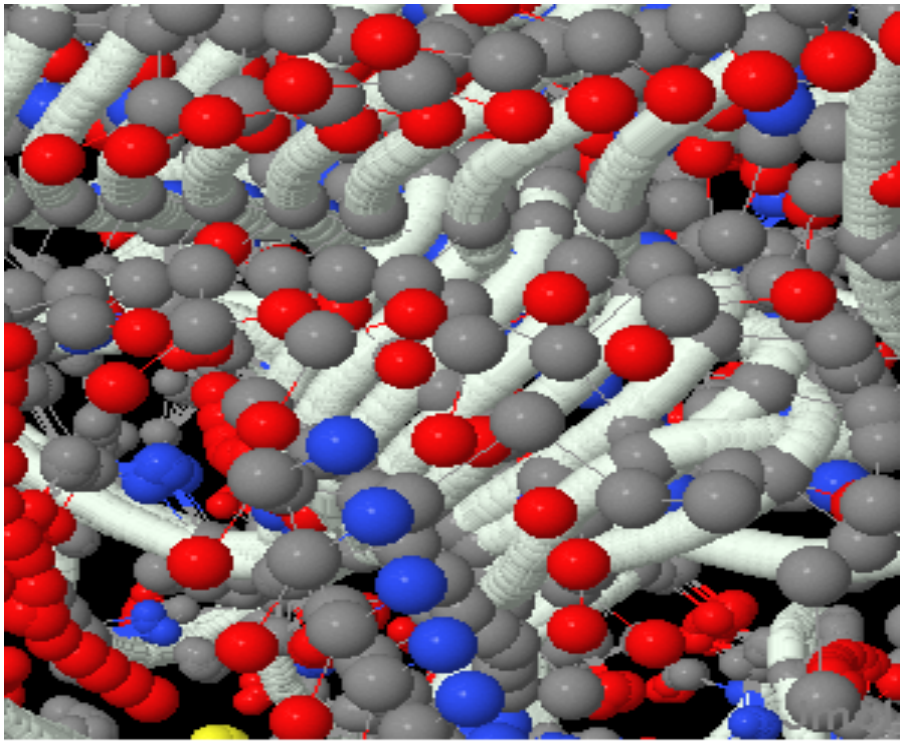


Figura 20. Imagem demonstrando rastros que as moléculas deixaram.

5.2. CLUSTALW – ALINHAMENTO DE SEQUENCIAS

CLUSTALW é a ferramenta responsável por realizar os alinhamentos de sequencias biológico, passo obrigatório para o avanço de pesquisas como citado em capítulos anteriores.

Trata-se da ferramenta mais popular na implementação de múltiplos alinhamentos de sequencias de DNA.

Desenvolvido pelo Dr. *Kuo-Bin Li* do instituto de bioinformática de Singapura, é um software que embora possua características de software livre, como permissão de execução de programa, de estudo de seu funcionamento, de liberdade de redistribuição e de modificação, não há permissão de uso para fins comerciais. Realiza a leitura e gravação em arquivos com extensão PDB disponibilizadas no wwPDB

5.3. BLAST – SIMILARIDADE DE SEQUÊNCIAS

Elaborado por *Eugene Myers, Stephen Altschul, Warren Gish, David J. Lipman e Webb Miller* no *National Institutes of Health*, BLAST, derivado do Termo *Basic Local Alignment Search Tool* é um algoritmo que realiza consultas em banco de dados por sequencias similares a uma sequência-alvo. O BLAST e suas variáveis verificam cada entrada no banco de dados de forma independente contra a sequência-alvo.

As diversas variáveis do BLAST são citadas na tabela a seguir.

Programa BLAST e suas Variáveis		
Programa	**Tipo de Sequencia-alvo**	**Consulta em Banco de Dados de**
BLASTP	Sequências de aminoácidos	Sequências de proteínas
BLASTX	Sequências de nucleotídeos traduzidas	Sequências de proteínas
TBLASTN	Sequências de aminoácidos	Sequências de nucleotídeos traduzidas
TBLASTX	Sequências de nucleotídeos traduzidas	Sequências de nucleotídeos traduzidas
PSI-BLAST	Sequências de aminoácidos	Sequências de proteínas

Tabela 1. Programa BLAST e suas variáveis

Para entendimento do que a aplicação faz podemos citar como exemplo a descoberta de um novo gene de camundongo. Após essa descoberta realizamos uma pesquisa no BLAST do genoma humano a fim de verificar se existem seres humanos que portam genes semelhantes.

5.4. ARTEMIS – GERADOR DE ANOTAÇÕES

O ARTEMIS é uma ferramenta de anotação que permite a visualização de anotações de sequência e os resultados de análise com o contexto da sequência.

Desenvolvido pelo *The Sanger Institute*, o ARTEMIS é um software livre que está disponível para qualquer sistema operacional em uma versão desenvolvida em Java.

Executa arquivos na extensão FASTA, que se trata de arquivos que possuem representações de sequencias genéticas. (BELOZZE, 2007)

Dentre os tipos citados de anotações, a principal gerada pelo aplicativo é a automática, porém podem ser complementadas com informações adicionais.

Na imagem abaixo é demonstrado um exemplo de anotações genômicas.

A mensagem destacada pela seleção em vermelho refere a uma anotação automática gerada pelo ARTEMIS demonstrando que da sequencia 14914 até 16068 há uma similaridade de seqüência.

Figura 21. Sequências e anotações geradas pelo software ARTEMIS

5.5. APOLLO – GERADOR DE ANOTAÇÕES

Outro sistema gerador automático e editor de anotações que pode ser citado é o APOLLO, de código aberto. Também permite que os pesquisadores explorem anotações em vários níveis de detalhe e criem anotações manuais, tudo em um ambiente gráfico, contudo o sistema não suporta anotação baseada em ontologia. (BELOZZE, 2007)

A aplicação também disponibiliza para execução diversas ferramentas de análise de dados genômicos como GENSCAN e BLAST (busca por similaridade entre as sequências) (BRITO, 2003).

Os milhares de sequências são importadas e gravadas através de um arquivo XML com padrão definido.

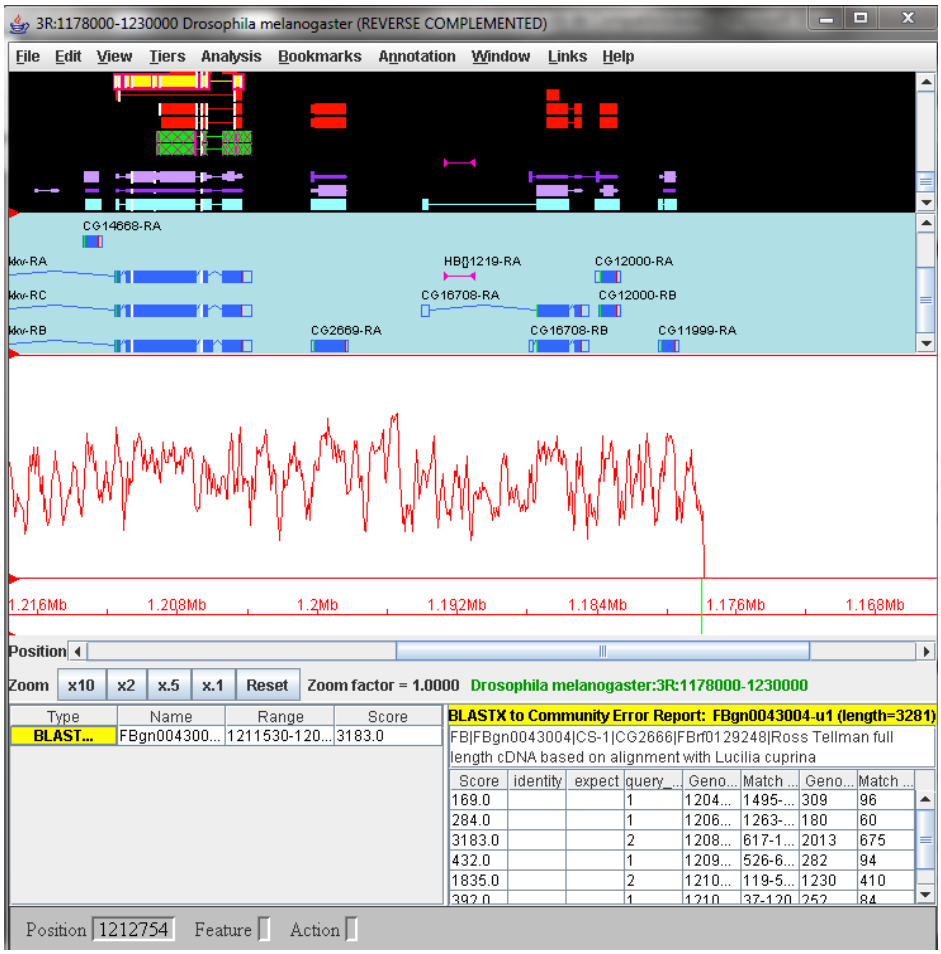


Figura 22. Resultado após analise de arquivo XML

Na imagem abaixo é mostrado uma tela de alinhamento de sequência, passo necessário para o levantamento automático de anotações. A operação de alinhamento de sequencia é o detalhe necessário para avanço de pesquisas estudado no capitulo anterior.

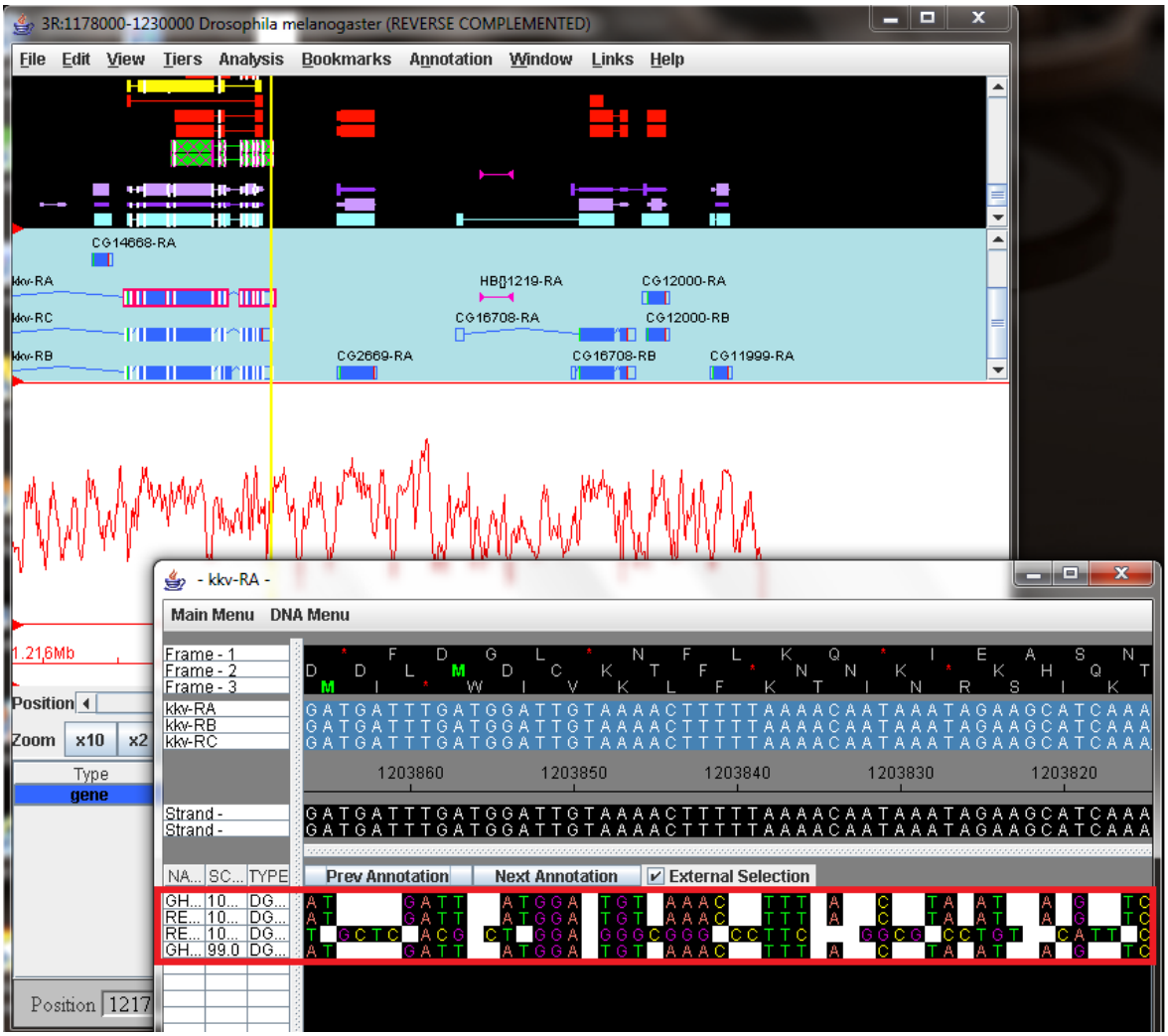


Figura 23. Alinhamento de sequência realizado pelo Apollo

Na imagem a seguir é demonstrado um trecho nas quais essas sequências são escritas em XML para que a aplicação possa interpretar.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- DOCTYPE game SYSTEM "GAME.dtd" -->
<game>
  <!-- Curational annotations from Apollo -->
  <!-- Analysis of: 3R:1178000-1230000 -->
  <!-- Saved on Tue May 10 17:35:09 GMT 2005 -->
  <!-- Apollo version: Apollo Genome Annotation and Curation Tool
/users/nomi/apollo/conf/apollo.cfg -->
  <seq id="3R:1178000-1230000" length="52001" focus="true">
    <name>3R:1178000-1230000</name>
    <organism>Drosophila melanogaster</organism>
    <potential_sequencing_error>
      <type>deletion</type>
      <position>1222356</position>
    </potential_sequencing_error>
    <residues>
      AAGCCCACTATATTGCATTAAATTATGCGATAATTGATCAATTTTAAAGG
  <!--
        1040 linhas de sequencias foram excluidas do exemplo
    -->

      ATTATATATAATGTTAAGACGTTCATTTCTGTAGCAATATAAGCAAAATG
      CGATACTATGTTCTGGTAATATGACATCGTGGACAGCTATGTGCTATCCA
    </residues>
  </seq>
  <map_position type="tile" seq="3R:1178000-1230000">
    <chromosome>3R</chromosome>
    <organism>Drosophila melanogaster</organism>
    <arm>3R</arm>
    <span>
      <start>1178000</start>
      <end>1230000</end>
    </span>
  </map_position>
  <annotation id="CG1161">
```

Figura 24. Exemplo de arquivo XML para análise pelo software Apollo

6. CONSIDERAÇÕES FINAIS E PROJEÇÕES FUTURAS

Neste trabalho foram levantados e apresentados os variados tipos de dados biológicos, banco de dados em si e as aplicações mais utilizadas atualmente para o gerenciamento de informações biológicas.

Para os diversos tipos de dados cada um deve ser armazenado em um determinado padrão. Devido essa diversidade de tipos existentes e sistemas de bancos de dados disponíveis, não há uma forma de se definir um modelo relacional padrão para apenas um Banco de Dados.

A vantagem observada é que mesmo com as diversidades de tipos levantadas é que não tem sido necessário a criação de novas bases com novos formatos de armazenamento para um novo dado descoberto por pesquisadores, porém as informações são limitadas para as determinadas aplicações.

A desvantagem identificada é quanto à integração dos dados entre as aplicações citadas. Há um alto número de aplicações que executam sobre uma base de dados específica, com exceção dos que leem e gravam em bases PDB disponibilizadas. Há as que possuem certo nível de integração, porém não possuem mecanismos disponíveis para realizar todas as validações necessárias, como é o caso dos dados compartilhados entre os laboratórios responsáveis pelas bases DataBank, DDBJ e EMBL.

A atualização por meio de arquivos de textos enviados por correio eletrônico geram desconfortos, já que nas maiorias das vezes ocorre rejeição por inconsistência de alguma entrada.

Para um trabalho futuro, um Banco de Dados Relacional que seja flexível à criação de novos tipos de dados poderá ser criado. Com isso poderá organizar tudo o que se tem de pesquisas levantadas, criando um banco de dados central com todas as informações concisas integradas entre os distintos laboratórios e garantir que não ocorra perda de informações. Para essa base é muito importante que exista um algoritmo que verifique cada sequencia enviada com os dados já armazenados para o fato da resolução também da redundância dos dados.

REFERÊNCIAS

LIFSCHITZ, Sérgio. **"Gerenciadores de Dados Biológicos: Genéricos ou Ad-hoc?",** Departamento de Informática Pontifícia Universidade Católica do Rio de Janeiro, 2007

BANERJEE, S. **"A Database Platform for Bioinformatics"**, Oracle Corporation, Redwood Shores, 2000.

OLIVEIRA, Thiago Y.K. **"Introdução a Bioinformática e Banco de Dados Biológicos",** Departamento de Genética/ FMRP/USP, 2010

RIDLEY, M. **"Genoma"**, Record, 2001.

BERMAN, Helen M. "**Rutgers Chemistry and Chemical Biology Department**"

LESK, Arthur M. "**Introdução à Bioinformática",** 2.ed. Porto Alegre Artmed, 2008

PDB - Protein Data Bank, http://www.icb.ufmg.br/prodabi/tour/pdb.html, acessado em outubro de 2010

wwPDB - Protein Data Bank, http://www.rcsb.org/ acessado em setembro de 2010

BRITO, Rogério T. **"Dissertação apresentada ao Instituto de Matemática e Estatística Da Universidade de São Paulo para obtenção do grau de Mestre em Ciência da Computação",** Universidade de São Paulo, 2003

LEHNINGER, Albert L. "***Princípios de Bioquímica***", 1995. 839 p.

BELLOZE, Kelli T., "**Uma extensão do processo de anotação genômica para ampliar o uso e a evolução colaborativa de ontologias no domínio da biologia molecular**", Instituto Militar de Engenharia do Rio de Janeiro, 2007.

SOARES, Paulo R., "Introdução à Biologia Molecular Computacional (IF803)", UFPE, 2010

DDBJ – DNA Data Bank of Japan, http://www.ddbj.nig.ac.jp, acessado em Outubro de 2010

The EMBL Nucleotide Sequence Database, http://www.ebi.ac.uk/embl/, acessado em Outubro de 2010