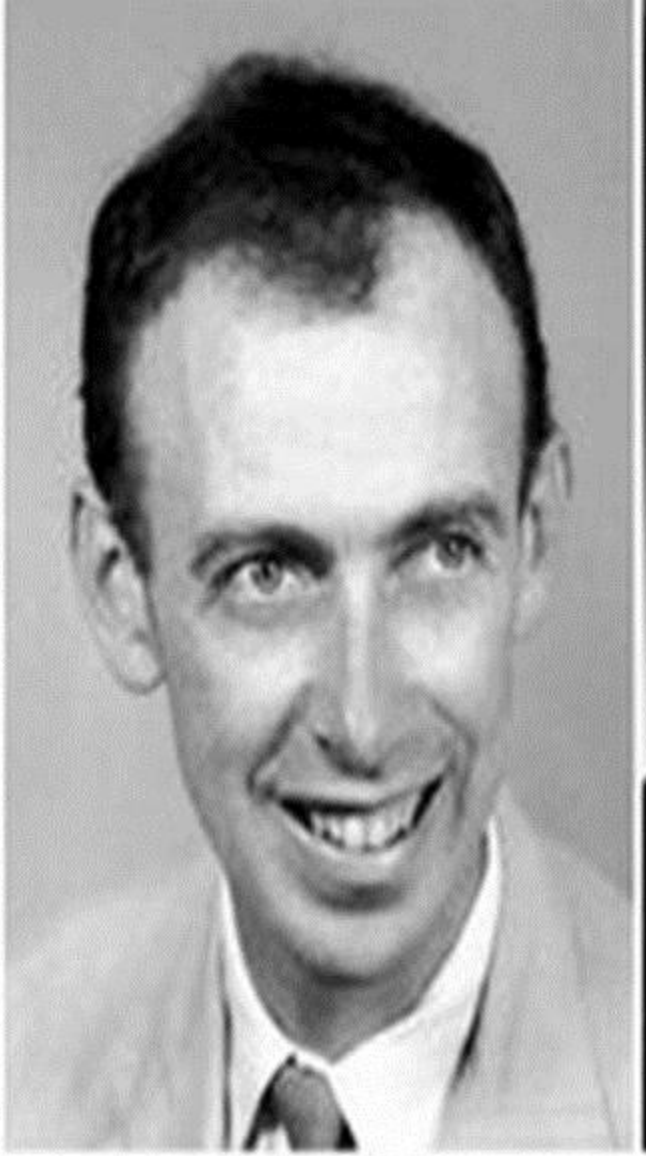
Bioinformática Aplicada ao Melhoramento Vegetal e Microrganismos

Tiago Benedito dos Santos

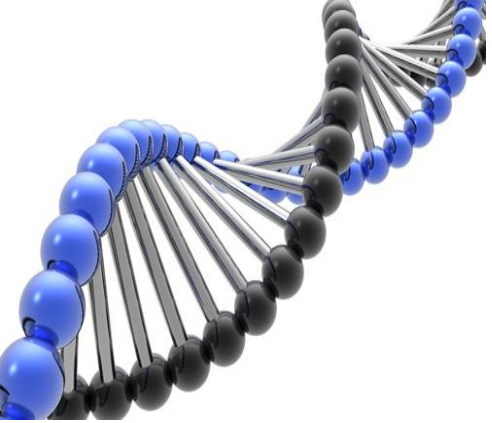
Francis Harry
Compton Crick
(1916-2004)



James Dewey
Watson
(1928 -)

Maurice Hugh
Frederick Wilkins
(1916-2004)

Proposta da Dupla hélice de DNA

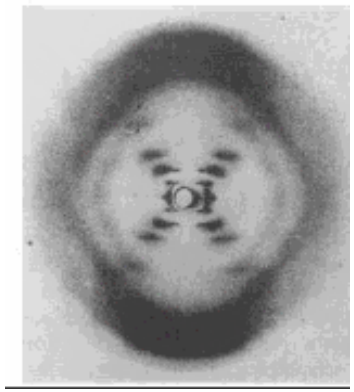


DNA, o código da vida...

- *Molecular Structure of Nucleic Acids. April 25, 1953*
- James D. **Watson** e Francis **Crick**


April 25, 1953 NATURE

MOLECULAR STRUCTURE OF NUCLEIC ACIDS




Prêmio Nobel 1962
Wilkins
Perutz
Crick
Steinbeck
Watson
Kendrew

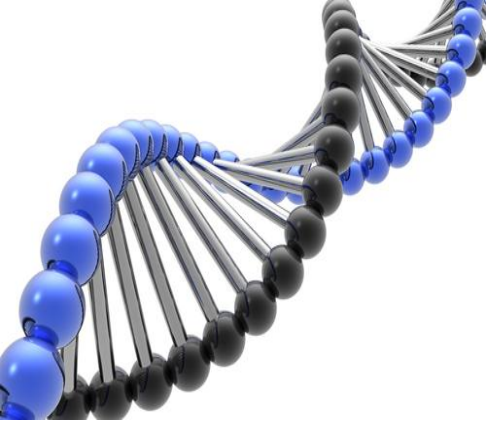
Contribuições para o modelo :

Erwin Chargaff: **Quantificou as bases nitrogenadas adenina, timina, citosina e guanina em diferentes organismos e demonstrou existir uma razão constante entre A/T e C/G**
Evidência para pareamento entre as bases nas duas fitas de DNA

Rosalind Franklin: **Paralelamente a Watson e Crick, realizou estudos de difração de raios-X com DNA.**
Seus dados foram fundamentais à proposição da estrutura da dupla hélice.

Linus Pauling: **Descreveu a estrutura de hélice para proteínas.**
Mostrou que estrutura de hélice era possível.

Frederick Sanger: **Desenvolveu o método de sequenciamento de proteínas e relatou a sequência da insulina.**
Evidência de que proteínas não eram polímeros simples, dependendo necessáriamente de um código.



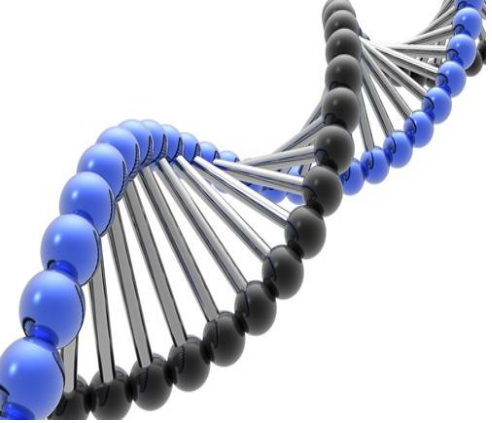
A Verdadeira Revolução

Início do séc. 20: Mendel e as leis da hereditariedade.

1953: Watson/Crick e a estrutura do DNA.

Anos 70 e 80: Biologia Molecular/Biotecnologia

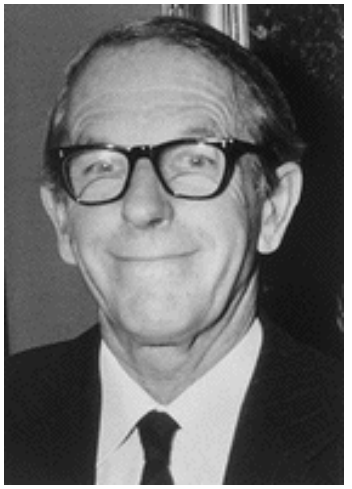
Anos 90 e séc. 21: Genômica/Bioinformática



História da Biologia Molecular

- **1951** Fred Sanger, Amino Acid Sequence of Insulin
- **1953** Watson/Crick, Estrutura do DNA
- **1957** Francis Crick, Central Dogma, DNA → RNA → Protein
- **1960's** Nirenberg, Matthaei, The Genetic Code
- **1967** Shapiro and Beckwith, First gene cloned, LacZ
- **1972** Paul Berg, First recombinant DNA molecule
- **1973** Cohen/Boyer, First recombinant organism
- **1977** Maxam/Gilbert and Fred Sanger, DNA sequencing
- **1977** Fred Sanger, Complete sequence of phage ϕX174
- **1978** David Botstein, Restriction Fragment Length Polymorphisms (RFLP)
- **1980** Kerry Mullis, PCR

Frederick Sanger 1918-2013



A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase

F. SANGER AND A. R. COULSON

Medical Research Council
Laboratory of Molecular Biology
Hills Road, Cambridge CB2 2QH, England

Prêmio Nobel de medicina e fisiologia em 1980

J. Mol. Biol. v.94, p. 441-448, 1975

1986 - Sequenciador Automático de DNA (Leroy Hood)

Lloyd M. Smith et al. Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-679 (June 12, 1986).

T. Hunkapiller, R. J. Kaiser, B. F. Koop, L. Hood. Large-Scale and automated DNA sequence determination. *Science* **254**. 59-67 (October 4. 1991).

Sequenciamento de DNA

- 1977 -2004 – Evolução e consolidação do método de Sanger
- 2004 - ……. – Plataformas de NGS – Next Generation Sequencing (Na verdade pode ser: “Now Generation Sequencing”)

Gerações de Sequenciadores

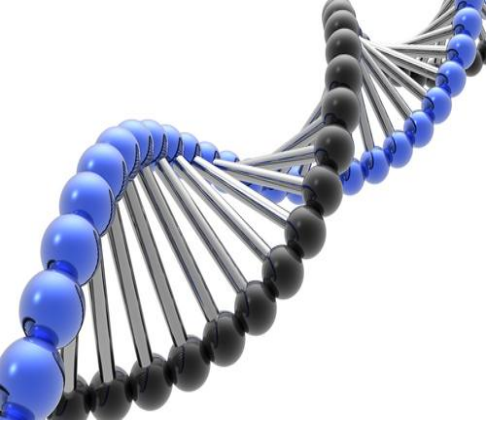
Sequenciamento – nova geração

Applied Biosystems
ABI 3730XL
1 Mb / dia

Roche / 454
Genome Sequencer FLX
100 Mb / corrida

Applied Biosystems
SOLiD
3000 Mb / corrida

Illumina / Solexa
Genetic Analyzer
2000 Mb / corrida



Sequenciando Genomas para Investigar Variabilidade

- SNVs – Single nucleotide variants
- SNPs – Single nucleotide polymorphisms
- CNVs – Copy number variants
- CNPs – Copy number polymorphisms

Genomas Individuais

OPEN ACCESS Freely available online October 2007 | Volume 5 | Issue 10 | e254 PLOS BIOLOGY

The Diploid Genome Sequence of an Individual Human

Samuel Levy[1*], Granger Sutton[1], Pauline C. Ng[1], Lars Feuk[2], Aaron L. Halpern[1], Brian P. Walenz[1], Nelson Axelrod[1], Jiaqi Huang[1], Ewen F. Kirkness[1], Gennady Denisov[1], Yuan Lin[1], Jeffrey R. MacDonald[2], Andy Wing Chun Pang[2], Mary Shago[2], Timothy B. Stockwell[1], Alexia Tsiamouri[1], Vineet Bafna[3], Vikas Bansal[3], Saul A. Kravitz[1], Dana A. Busam[1], Karen Y. Beeson[1], Tina C. McIntosh[1], Karin A. Remington[1], Josep F. Abril[4], John Gill[1], Jon Borman[1], Yu-Hui Rogers[1], Marvin E. Frazier[1], Stephen W. Scherer[2], Robert L. Strausberg[1], J. Craig Venter[1]

http://huref.jcvi.org/

J. Craig Venter™ INSTITUTE

Table 4. Identification of Variants Found within the HuRef-NCBI One-to-One Assembly Map (Internal HuRef-NCBI map) and Those Variants in HuRef Sequence Not Aligned to NCBI (External HuRef-NCBI Map)

Variant	Internal HuRef-NCBI Map	External HuRef-NCBI Map
heterozygous SNP	1,623,826	138,715
homozygous SNP	1,450,860	—
heterozygous MNP	11,825	27,160
homozygous MNP	14,838	—
heterozygous indel	218,301	45,622
complex	5,880	22,299
homozygous insertion	—	275,512
homozygous deletion	—	283,961
inversion	—	90
Total	3,325,530	793,359

nature Vol 452|17 April 2008|doi:10.1038/nature06884

LETTERS

The complete genome of an individual by massively parallel DNA sequencing

David A. Wheeler[1]*, Maithreyan Srinivasan[2]*, Michael Egholm[2]*, Yufeng Shen[1]*, Lei Chen[1], Amy McGuire[3], Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[2]†, Faheem Niazi[2], Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[4], Xing-zhi Song[1], Yue Liu[1], Ye Yuan[1], Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinstock[1,4], Richard A. Gibbs[1,4] & Jonathan M. Rothberg[2]†

http://jimwatsonsequence.cshl.edu/cgi-perl/gbrowse/jwsequence/

Table 1 | Single nucleotide variation in 454 reads

Subject	Filter*	Total variation	Known†	Novel
Watson	Raw	14,829,087	3,283,273	11,545,814
	1	4,427,488	2,815,322	1,612,166
	2	3,971,513	2,752,991	1,218,522
	3	3,322,093	2,715,296	606,797
Venter‡	4	3,470,669	2,822,902	647,767

* Filters: raw, all base substitution from cross_match alignments; 1, $S_v > 28$ (see Methods); 2, filter 1 plus ratio of variant to total coverage >0.2; 3, filter 2 plus eliminate SNPs close to homopolymer runs >5bp; 4, (Venter) Phred (ref. 20) $Q > 15$, ratio of variant to total coverage >0.2.
† Variants found in build 126 of dbSNP (http://www.ncbi.nlm.nih.gov/SNP/).
‡ SNPs found in genome of Venter: see ref. 2 and supplementary material therein.

Genomas Individuais

Vol 456|6 November 2008|doi:10.1038/nature07517 nature

ARTICLES

Accurate whole human genome sequencing using reversible terminator chemistry Bentley et al.

http://www.illumina.com/HumanGenome/

nature Vol 456|6 November 2008|doi:10.1038/nature07484

ARTICLES

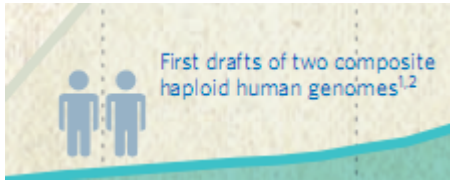
The diploid genome sequence of an Asian individual Wang et al.

http://yh.genomics.org.cn/

华大基因 BGI SHENZHEN

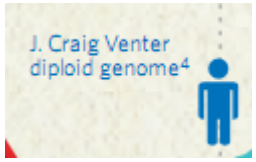
YanHuang
——The First Asian Diploid Genome

Explosão de Genomas Individuais



2001

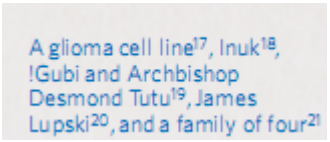
2004

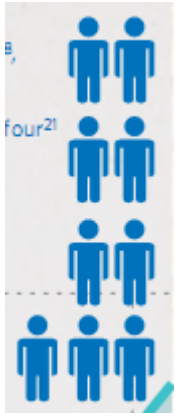


2007

2008

2009





2010

Mais e mais Genomas Individuais

PRESS RELEASES AND MEDIA COVERAGE

Recent Media Coverage

IN SEQUENCE: TUESDAY JUNE 22, 2010
1000 Genomes Project Releases Pilot Project Data, Plans to Sequence 2,500 Genomes in Total

GENOME WEB DAILY NEWS: FRIDAY MAY 14, 2010
1000 Genomes Project to Sequence 2,500 Genomes By End of Next Year

GENOME WEB DAILY NEWS: THURSDAY OCTOBER 22, 2009
Researchers Discuss 1000 Genomes Progress at ASHG

Press Releases

MONDAY JUN. 21, 2010
1000 Genomes Project Releases Data from Pilot Projects on Path to Providing Database for 2,500 Human Genomes

WEDNESDAY JUN. 11, 2008
Three Sequencing Companies Join 1000 Genomes Project

TUESDAY JAN. 22, 2008
International Consortium Announces the 1000 Genomes Project

CURRENT ISSUE

Japanese Table of Contents

Volume 467 Number 7319 pp1005-1146 28 October 2010

About the cover

ARTICLE

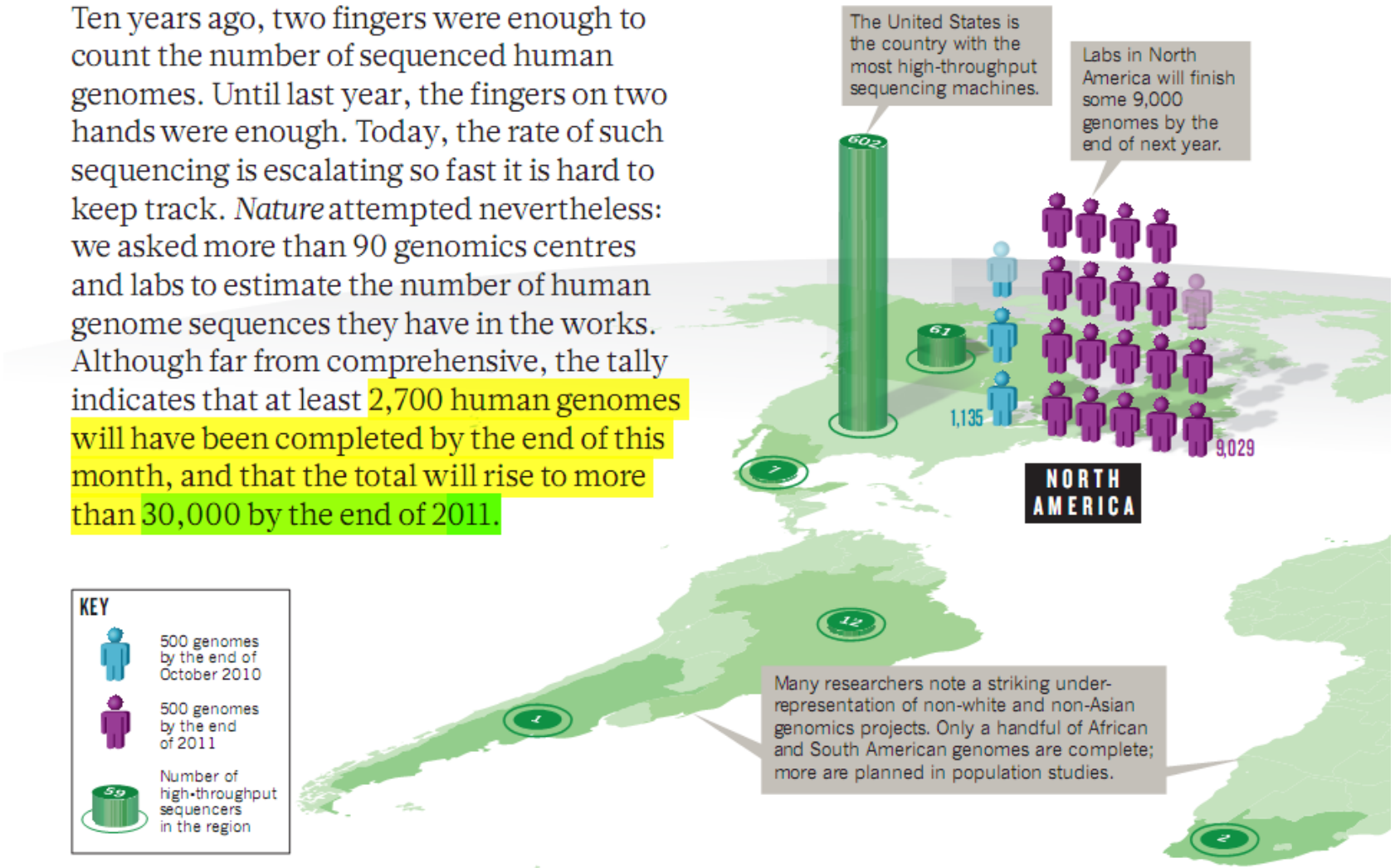
doi:10.1038/nature09534

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

Genomes by the thousand

Ten years ago, two fingers were enough to count the number of sequenced human genomes. Until last year, the fingers on two hands were enough. Today, the rate of such sequencing is escalating so fast it is hard to keep track. *Nature* attempted nevertheless: we asked more than 90 genomics centres and labs to estimate the number of human genome sequences they have in the works. Although far from comprehensive, the tally indicates that at least 2,700 human genomes will have been completed by the end of this month, and that the total will rise to more than 30,000 by the end of 2011.



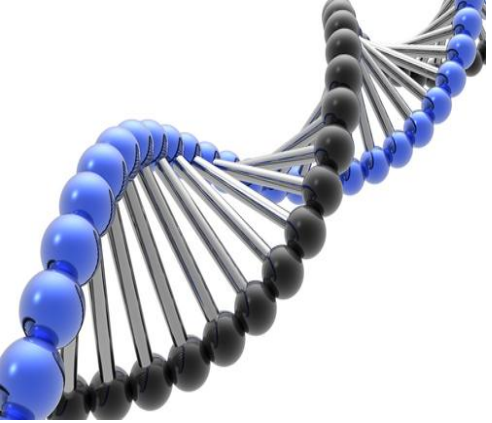
Antropologia Molecular

Kissing cousins. A few Neandertals mated with early modern humans and passed on some of their genes to living humans.

CREDIT: TOMISLAV MARICIC

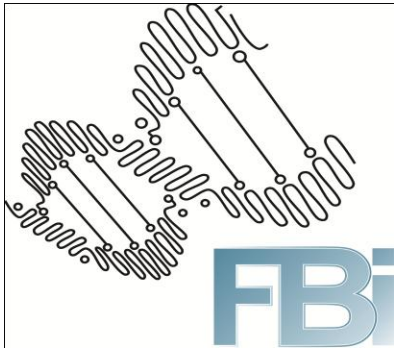
A Draft Sequence of the Neandertal Genome

Richard E. Green,[1*†‡] Johannes Krause,[1†§] Adrian W. Briggs,[1†§] Tomislav Maricic,[1†§] Udo Stenzel,[1†§] Martin Kircher,[1†§] Nick Patterson,[2†§] Heng Li,[2†] Weiwei Zhai,[3†‖] Markus Hsi-Yang Fritz,[4†] Nancy F. Hansen,[5†] Eric Y. Durand,[3†] Anna-Sapfo Malaspinas,[3†] Jeffrey D. Jensen,[6†] Tomas Marques-Bonet,[7,13†] Can Alkan,[7†] Kay Prüfer,[1†] Matthias Meyer,[1†] Hernán A. Burbano,[1†] Jeffrey M. Good,[1,8†] Rigo Schultz,[1] Ayinuer Aximu-Petri,[1] Anne Butthof,[1] Barbara Höber,[1] Barbara Höffner,[1] Madlen Siegemund,[1] Antje Weihmann,[1] Chad Nusbaum,[2] Eric S. Lander,[2] Carsten Russ,[2] Nathaniel Novod,[2] Jason Affourtit,[9] Michael Egholm,[9] Christine Verna,[21] Pavao Rudan,[10] Dejana Brajkovic,[11] Željko Kucan,[10] Ivan Gušic,[10] Vladimir B. Doronichev,[12] Liubov V. Golovanova,[12] Carles Lalueza-Fox,[13] Marco de la Rasilla,[14] Javier Fortea,[14¶] Antonio Rosas,[15] Ralf W. Schmitz,[16,17] Philip L. F. Johnson,[18†] Evan E. Eichler,[7†] Daniel Falush,[19†] Ewan Birney,[4†] James C. Mullikin,[5†] Montgomery Slatkin,[3†] Rasmus Nielsen,[3†] Janet Kelso,[1†] Michael Lachmann,[1†] David Reich,[2,20*†] Svante Pääbo[1*†]



Nova tecnologia

- Dispensa clonagem dos fragmentos em sistemas bacterianos
- Dispensa a preparação de DNA molde para sequenciamento
- Reações feitas em paralelo em volume extremamente pequeno - nanotecnologia



- ✓ **A primeira planta a ser sequenciada (*Arabidopsis thaliana*) levou aproximadamente 10 anos para ter o primeiro rascunho de seu genoma apresentado.**

- ✓ **utilização das segunda e terceira gerações de sequenciadores (*e.g.* 454, da Roche; SOLiD, da Applied Biosystems; *Genome Analyzer IIe,* da Illumina) poucas semanas.**

- ✓ **Quarta geração de sequenciadores (*e.g.* Oxford Nanopore, Pacific Biosciences, Ion Torrent) deve permitir, em poucas horas, ou até minutos, sequenciar o genoma completo de uma espécie a preços razoáveis que podem potencialmente chegar a menos de US$ 100/genoma.**

OXFORD NANOPORE TECHNOLOGIES LTD

Reader of genomes

Sequence DNA in seconds

DNA sequencing can now be done on a device that plugs into your computer like a memory stick

Duncan Graham-Rowe

IT LOOKS like an ordinary USB memory stick, but a little gadget that can sequence DNA while plugged into your laptop could

Genome Biology and Technology (AGBT) conference in Marco Island, Florida.

This was done as a proof of principle. "Phi X was the first genome ever to be sequenced," says the University of Birmingham, UK, and author of the blog Pathogens: Genes and Genomes. It shows the technology works, he says. "If you can sequence this genome you should be able to sequence larger genomes."

Oxford Nanopore is also building a larger device, called GridION, for lab use. Both GridION

to a solution containing enzymes that bind to the end of each strand. When a current is applied across the solution these enzymes and DNA are drawn to hundreds of wells in a membrane at the bottom of the solution, each just 10 micrometres across.

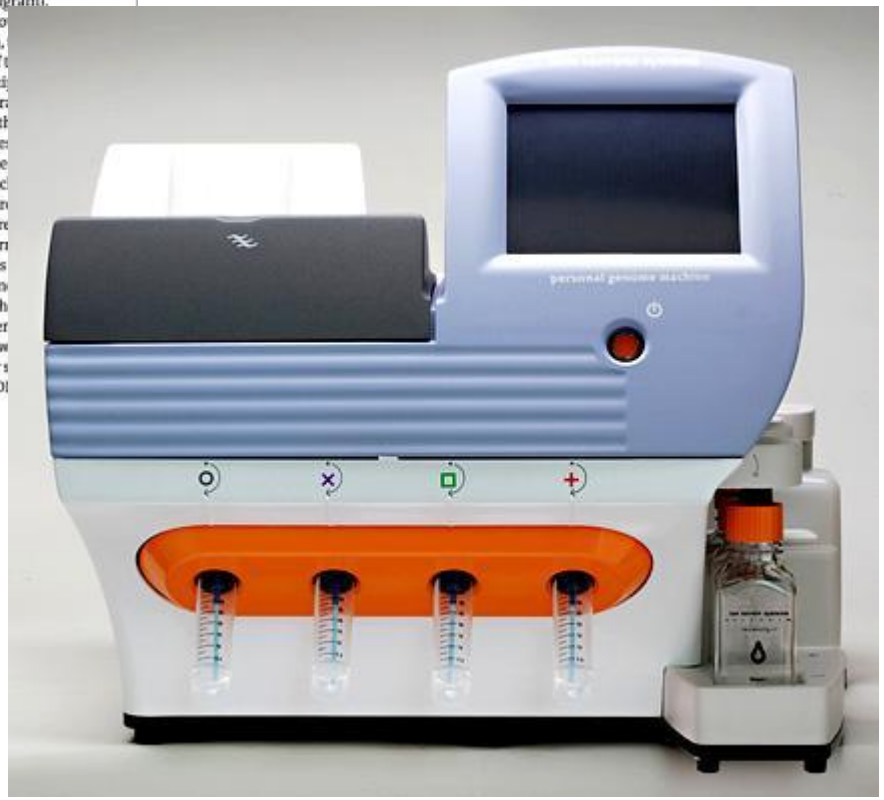
Within each well is a modified version of the protein alpha

"This could change the way we do medicine. You could see every doctor walking around with these things"

hemolysin (AHL), which has a hollow tube just 10 nanometres wide at its core (see diagram). As the DNA is drawn do towards to the protein, enzyme attaches itself t AHL and begins to unzi DNA, threading one str double helix through t

Each of the four base DNA strand has unique characteristics, and eac the current flowing thr pore in a slightly differe This is enough to deter which of the four bases through at any one tim disruption is read by th like a ticker-tape reader

This approach has tw advantages over other techniques: first, the D



Cost per Genome

$100M
$10M
$1M
$100K
$10K
$1K

2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013

Moore's Law

2001: *Human Genome Project*
2.7G$, 11 years

2001: *Celera*
100M$, 3 years

2007: *454*
1M$, 3 months

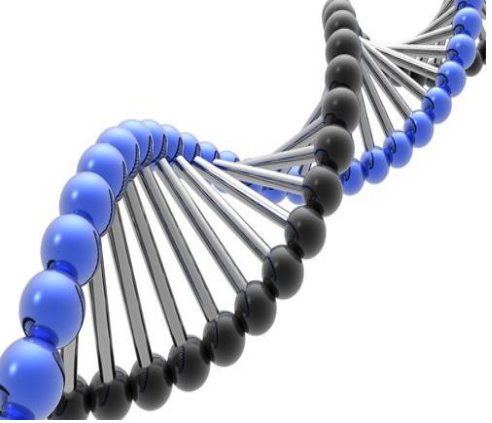
2008: *ABI SOLiD*
60K$, 2 weeks

2009: *Illumina, Helicos*
40-50K$

2010: 5K$, a few days?

2012: 1000$, <24 hrs?

NIH National Human Genome Research Institute
genome.gov/sequencingcosts

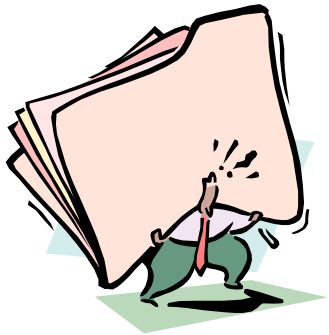


Aplicações

- Sequenciamento de Genomas
  - sequenciamento *de novo*
  - re-sequenciamento - variabilidade SNPs e mutações
- Sequenciamento de Transcriptomas
  - variabilidade - splicing, poliadenilação
  - quantificação de expressão gênica

Os sequenciadores de nova geração promovem uma mudança no paradigma

FIGURE 1



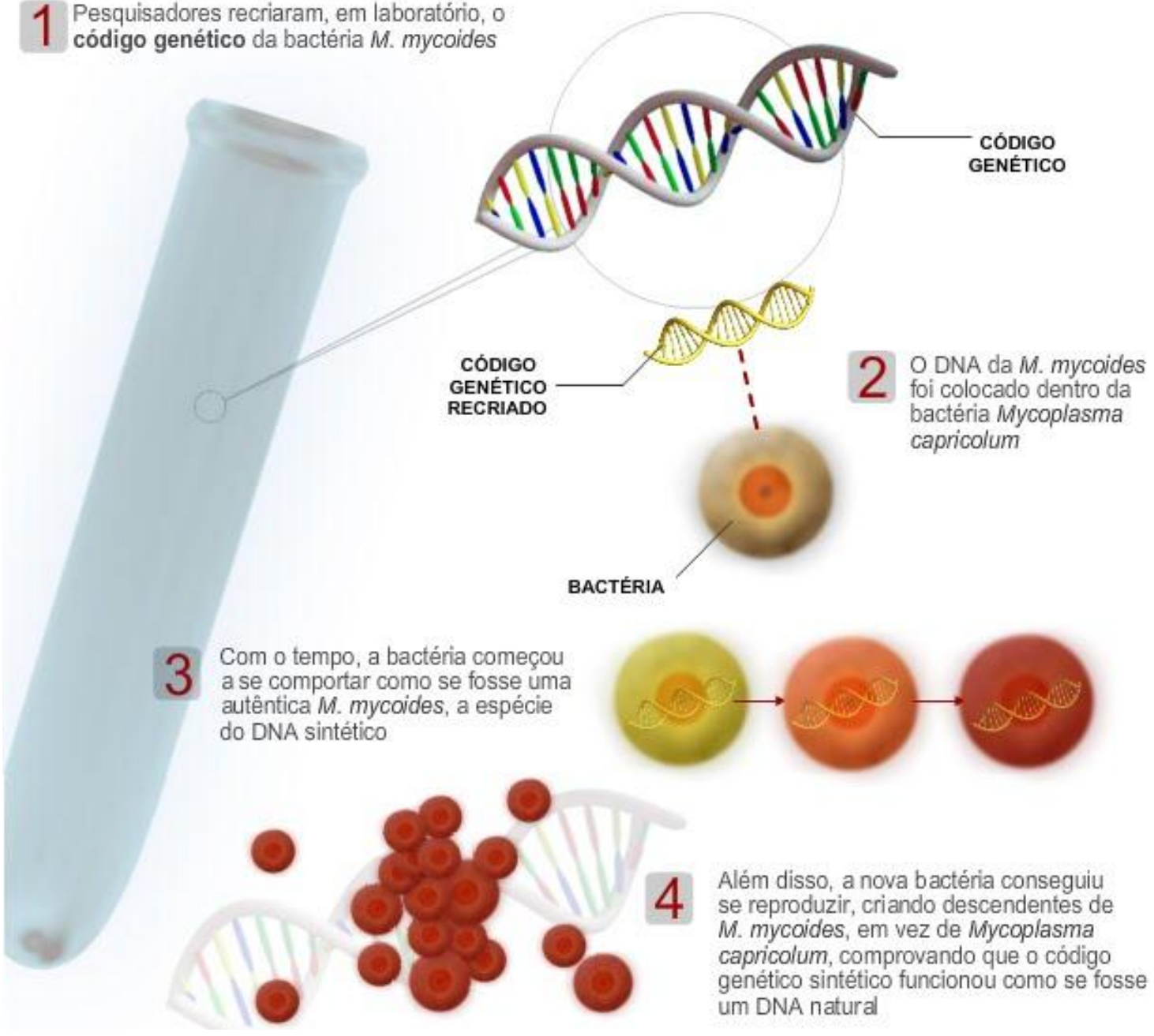
Geração de dados deixa de ser o fator limitante

Com os bilhões de datapoints gerados em horas, o processamento e análise dos dados tornou-se o maior gargalo das pesquisas.

Novas possibilidades

- ✓ Aprofundar regulação genoma
- ✓ Acompanhar resistência viral e bacteriana em tempo real
- ✓ Ter acesso a novos genes
- ✓ Entender a terapêutica gênica
- ✓ Integrar informações biológicas com questões de saúde e doença
- ✓ Avaliação genômica personalizada
- ✓ Avanços ainda inimagináveis

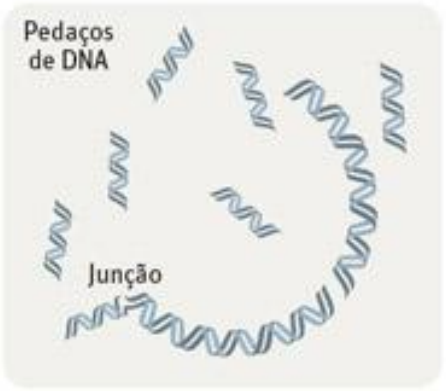
Entenda como foi feita a bactéria com DNA artificial



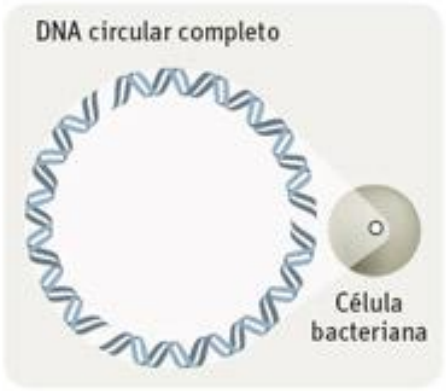
EDITORIA DE ARTE

SINTETIZANDO UM GENOMA FUNCIONAL

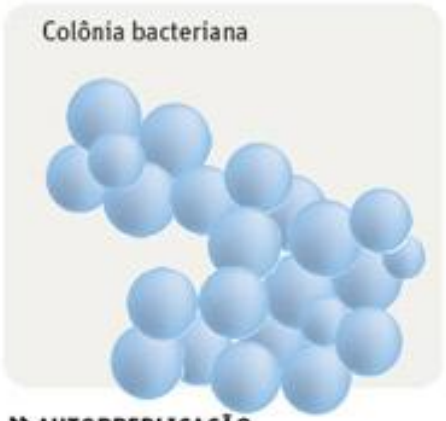
Uma equipe comandada por Craig Venter obteve sucesso na criação de um genoma bacteriano sintético e no seu uso para controlar uma célula.



» MONTAGEM
A equipe começou com pedaços pequenos de DNA feito em laboratório. Depois, ela usou uma técnica nova para juntá-los, criando o maior pedaço de DNA sintetizado da história, um círculo com um milhão de unidades de comprimento.



» INSERÇÃO
O DNA circular foi projetado para replicar quase fielmente a sequência genética da bactéria *Mycoplasma mycoides*. Para testar o DNA, a equipe o inseriu em uma célula vazia de uma espécie diferente de bactéria.



» AUTORREPLICAÇÃO
O DNA sintético se provou preciso o bastante para assumir a célula bacteriana e substituir o próprio DNA da célula. A "célula sintética" então se replicou para formar uma colônia bacteriana.

FONTE: REVISTA "SCIENCE"

J. Craig Venter Institute* (*JCVI*) publicaram resultados descrevendo o sucesso na construção da primeira bactéria sintética que se autorreplica. Eles sintetizaram os 1.8 milhões de pares de base cromossômicos do genoma modificado de *Mycoplasma mycoides.

O que é o GENOMA ????

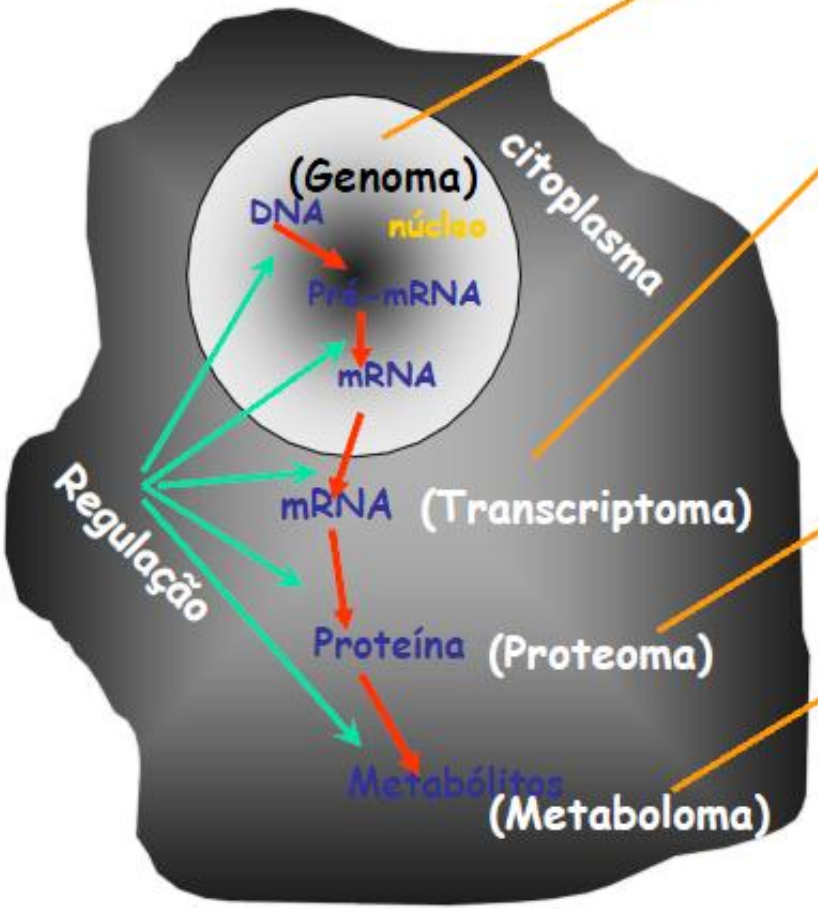
Todo o material genético contido nos cromossomos de um organismo

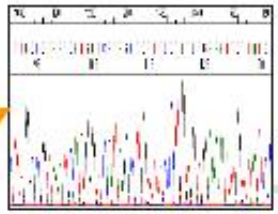
O genoma pode ser:

- **Estrutural** – genoma total do organismo

- **Funcional** – genoma relacionado aos genes expressos (EST's)

- Passou a ser "investigado" a partir das novas descobertas; e dos avanços tecnológicos na área da genética molecular.



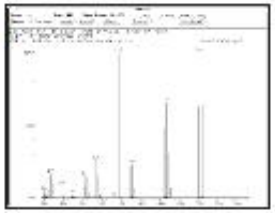


- Mapeamento
- Seqüênciamento
- Anotação

Genômica Estrutural

- Microarranjos de DNA
- Hibridização

- Eletroforese em Gel 2D
- Espectrometria de massa
- Seqüênciamento de proteínas



- Cromatografia
- Espectrometria

Genômica Funcional

Tipos de projeto

DNA – sequenciamento de estruturas do genoma ou de trechos destas. Ex.: Genoma humano

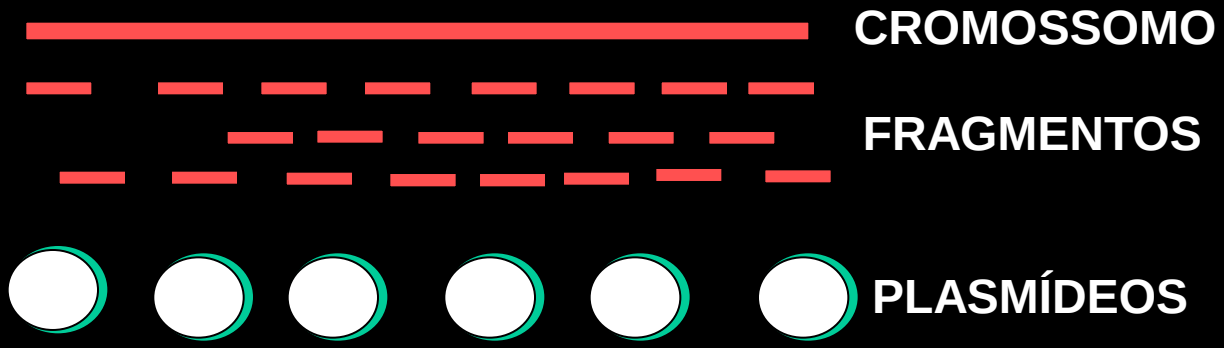
ESTs – sequenciamento de cDNA, feitos à partir de bibliotecas de mRNA. Ex.: ESTs de cana-de-açúcar

SAGE – sequenciamento de fragmentos em torno de 20 pb do cDNA

Objetivos para sequenciamento		
		Exemplo
Sequenciamento *de novo*	Sequenciamento genômico	Sequenciamento de >1000 genomas de influenza
	DNA de org. extinto	Neanderthal
	Metagenomica	Intestino humano
Resequencia-mento	Genomas completos	Indivíduos humanos
	Regiões genômicas	Detecção de rearranjos ou regiões associados à doenças
	Mutações somáticas	Em câncer
Transcriptoma	mRNA	Definir regulação da transcrição
	Serial Analysis of Gene Expression (SAGE)	
	RNAs não codificadores	Identificar e quantificar microRNAs
Epigenética	Padrão de Methilação	Avaliar padrão de metilação em câncer

Sequênciamento Genômico

- Nos permite conhecer as sequências nucleotídicas que compõem o organismo.
- Através deste conhecimento é possível a localização de genes de interesse.

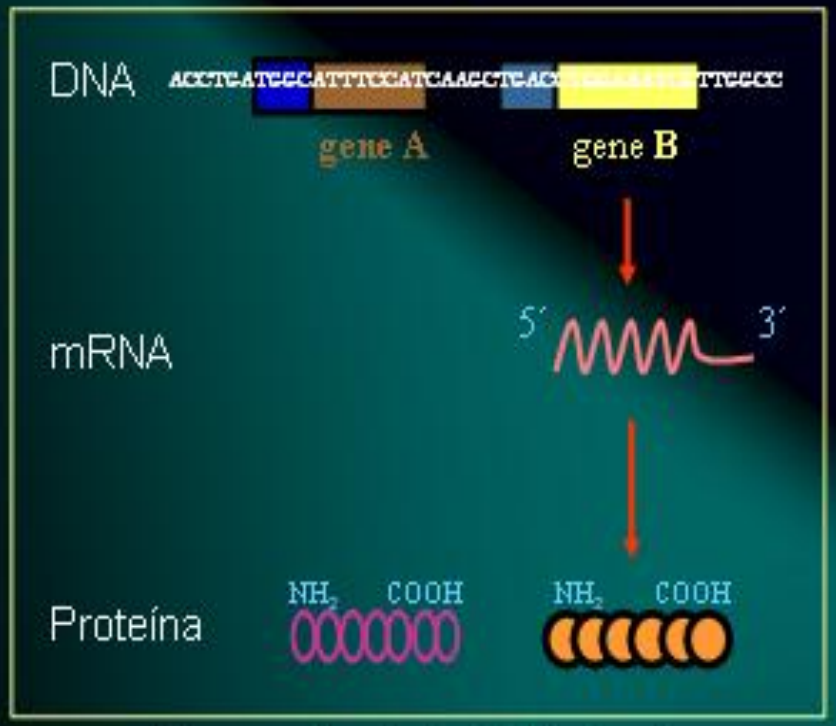


TRANSFORMAÇÃO BACTERIANA

MULTIPLICAÇÃO DE CLONES

EXTRAÇÃO DE DNA/mRNA

EST – *Expressed Sequence Tag*



Dogma Central da Biologia

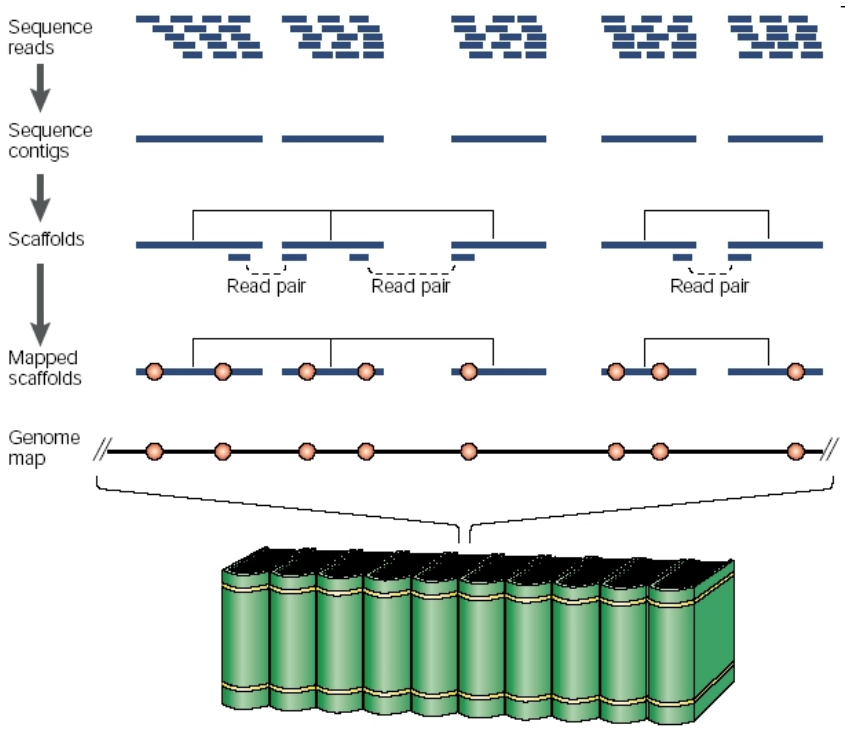
cDNA

ligação

Seqüenciamento ← Clonagem em *E. coli*

ESTRATÉGIAS DE SEQUENCIAMENTO DE GENOMAS

- Produção de *sequencing reads* até **8x** o tamanho do genoma
- Montagem dos **contigs**
- Identificação de **gaps** reais e virtuais
- Fechamento de gaps
- Publicação do genoma



HISTÓRICO PROJETOS GENOMAS

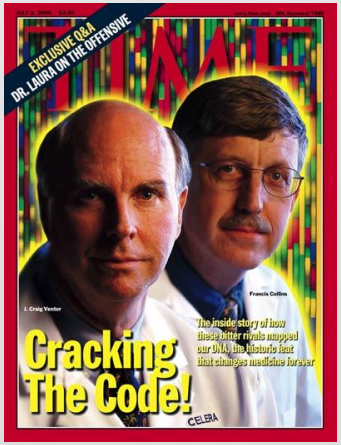
- **As primeiras propostas para se mapear o genoma surgiram em 1985, quando um grupo de cientista pretendiam detectar mutações em homens.**

- **1990 o Projeto Genoma Humano é iniciado formalmente.**

- **Basicamente, 18 países iniciaram programas de pesquisas sobre o genoma humano. Os maiores programas desenvolvem-se na Alemanha, Austrália, Brasil, Canadá, China, Coréia, Dinamarca, Estados Unidos, França, Holanda, Israel, Itália, Japão, México Reino Unido, Rússia, Suécia e União Européia.**

Francis Collins

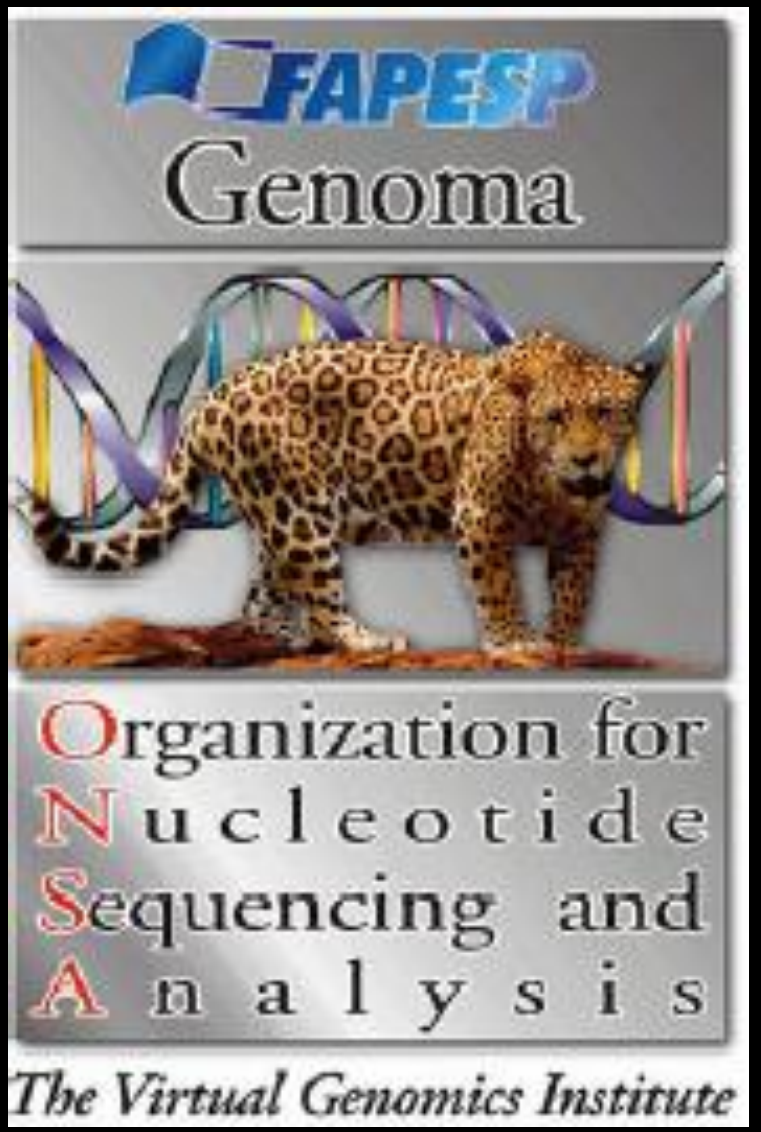
Craig Venter



No Brasil...

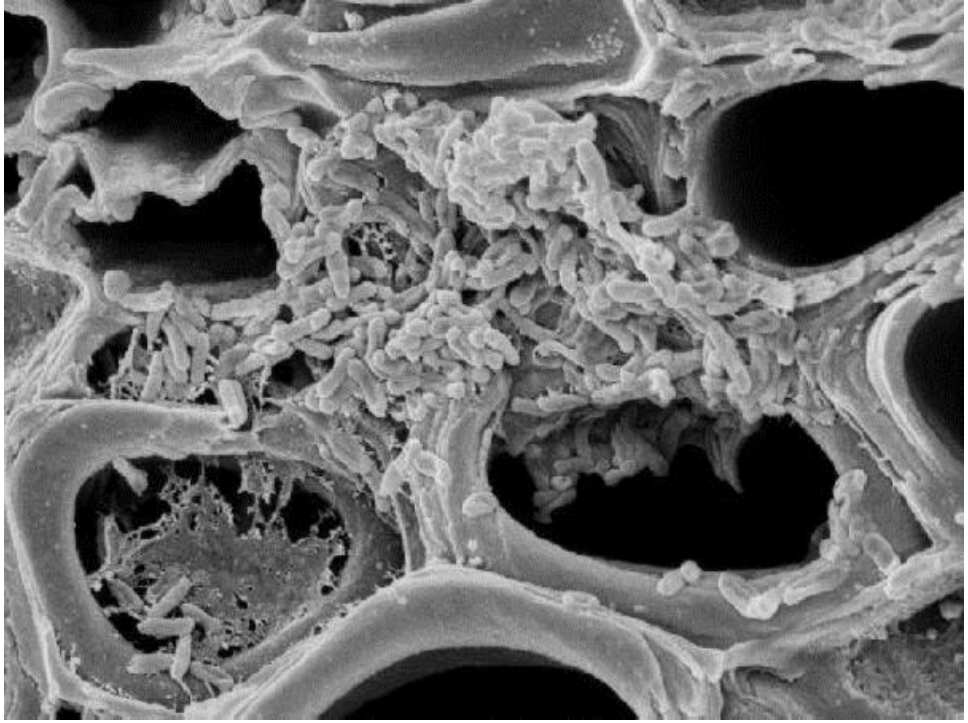
FAPESP 1997 - ONSA

Fundação de Amparo a Pesquisa do Estado de São Paulo



NOVEMBRO/99: SEQUENCIAMENTO, DESCOBERTA E DESCRIÇÃO DOS GENES FORAM CONCLUÍDOS

JULHO/2000: TRABALHO CIENTÍFICO FOI PUBLICADO NA REVISTA NATURE



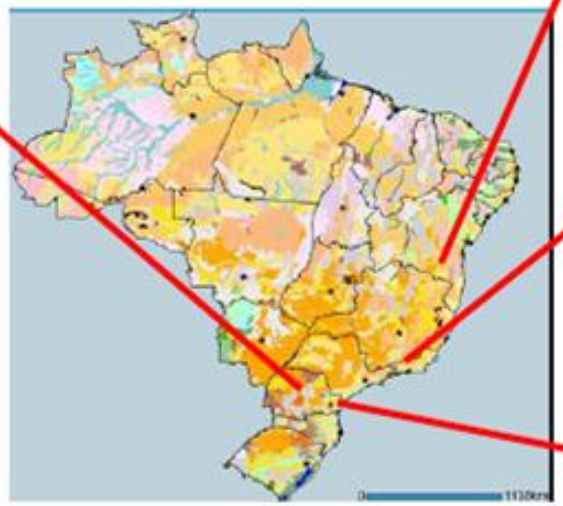

FAPESP
Genoma
Xylella

GENOMAS REGIONAIS BRASILEIROS

Rhizobium tropici

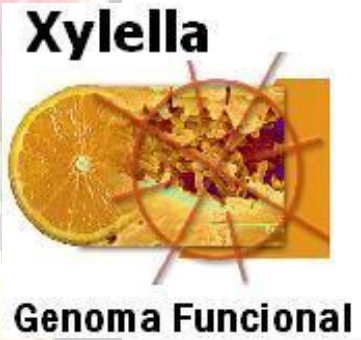
Área de Agricultura

Crinipellis perniciosa

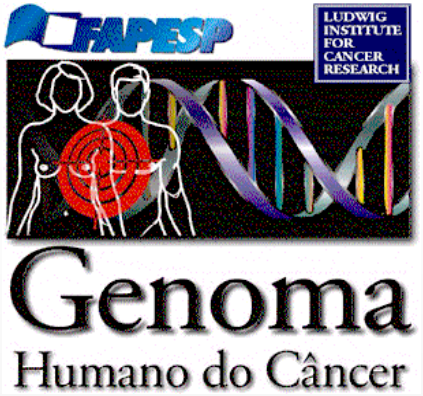


Herbaspirillum seropedicae

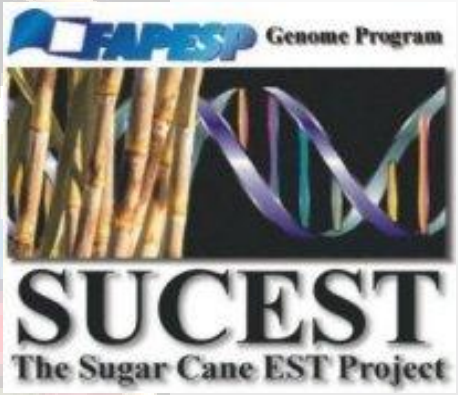
Gluconacetobacter diazotrophicus



Genoma Funcional da Xylella

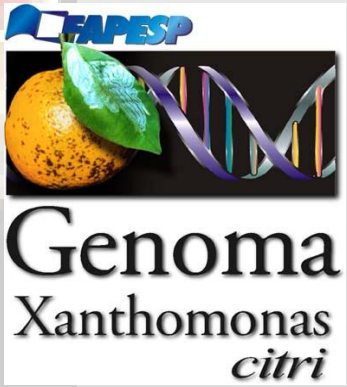


Genoma Humano do Câncer



Genoma da Cana de Açucar

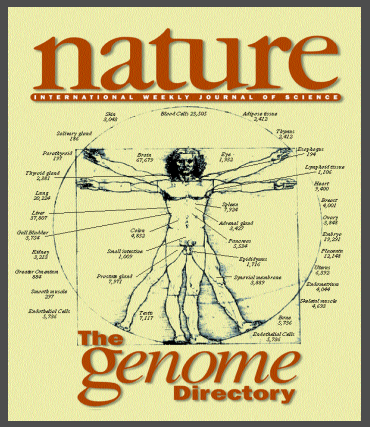
Genoma do *Schistosoma mansoni*

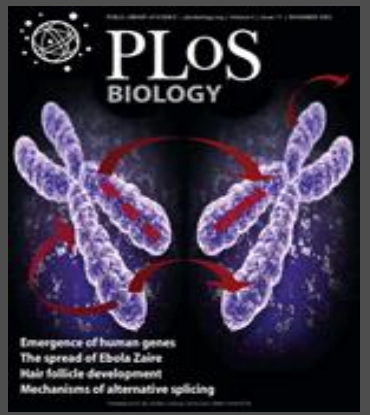


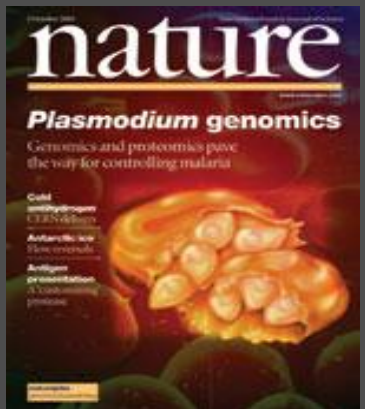
Genoma da *Xanthomonas citri* – cancro cítrico

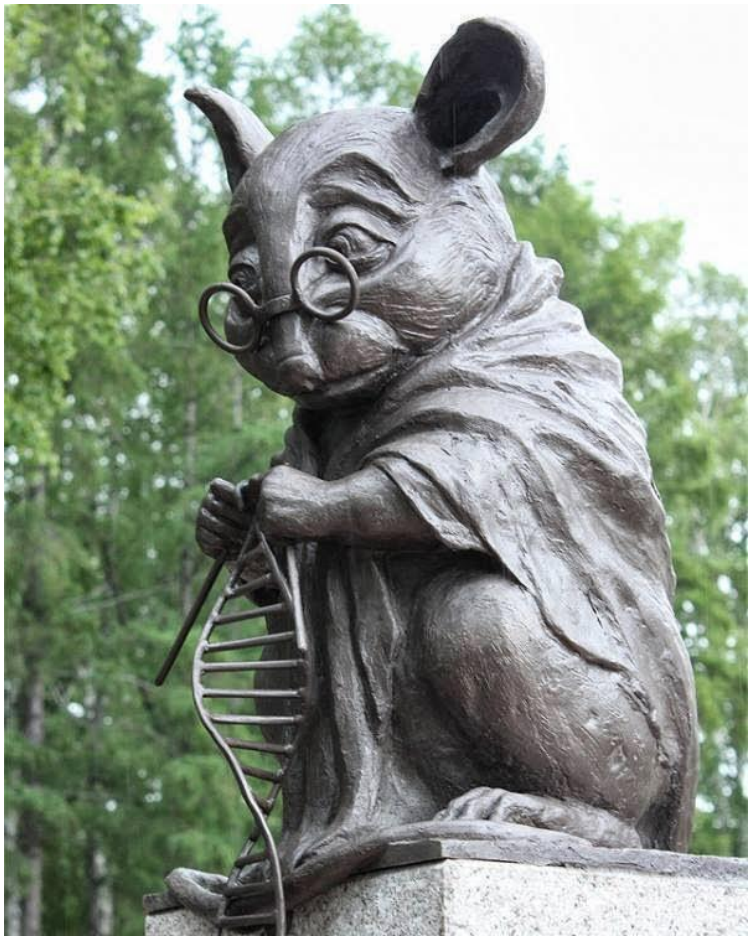
Genoma Café

Projetos Genomas já Publicados:









Monumento em homenagem aos ratos utilizados em pesquisas científicas sobre o DNA, em Novosibirski, na Rússia.


Homo sapiens


Arabidopsis thaliana


Mus musculus


C. elegans

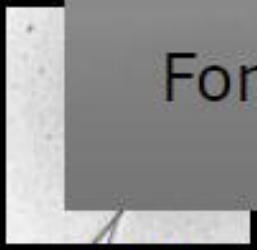

D. melanogaster

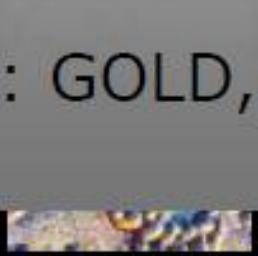
13.709 (11.192, 08/2011) Projetos Genoma
3.173 "completos" e/ou publicados

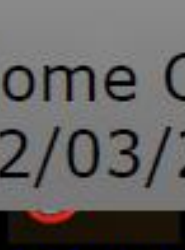
Em andamento:

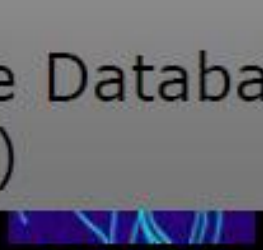
213 (206)	arqueobactérias
7.965 (6.806)	bactérias
2.385 (2.007)	eucariotos
334 (209)	metagenomas

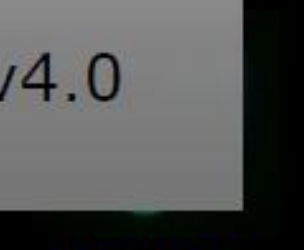
Fonte: GOLD, Genome Online Database v4.0 (12/03/2012)


E. coli


S. cerevisiae


Yersinia pestis


S. enterica


S. pombe

Tamanhos Comparativos de Alguns Genomas

Organismo	Número de pares de bases	Número de genes	Comentário
ΦX-174	5.386	10	vírus que infecta *E. coli*
Mitocôndria humana	16.569	37	organela subcelular
Vírus Epstein-Barr (EBV)	172.282	80	causador da mononucleose
Mycoplasma pneumoniae	816.394	680	causador de epidemias de pneumonia cíclica
Rickettsia prowazekii	1.111.523	834	bactéria, causadora de epidemias de tifo
Treponema pallidum	1.138.011	1.039	bactéria, causadora da sífilis
Borrelia burgdorferi	1.471.725	1.738	bactéria, causadora da doença de Lyme
Aquifex aeolicus	1.551.335	1.749	bactéria encontrada em fontes termais
Thermoplasma acidophillum	1.564.905	1.509	arqueobactéria sem parede celular
Campylobacter jejuni	1.641.481	1.708	causa freqüente de envenenamentos alimentares
Helicobacter pylori	1.667.867	1.589	principal causa de úlcera estomacal
Methanococcus jannaschii	1.664.970	1.783	arqueobactéria termófila
Haemophilus influenzae	1.830.138	1.738	bactéria, causadora de infecções do ouvido médio
Thermotoga maritima	1.860.725	1.879	bactéria marinha

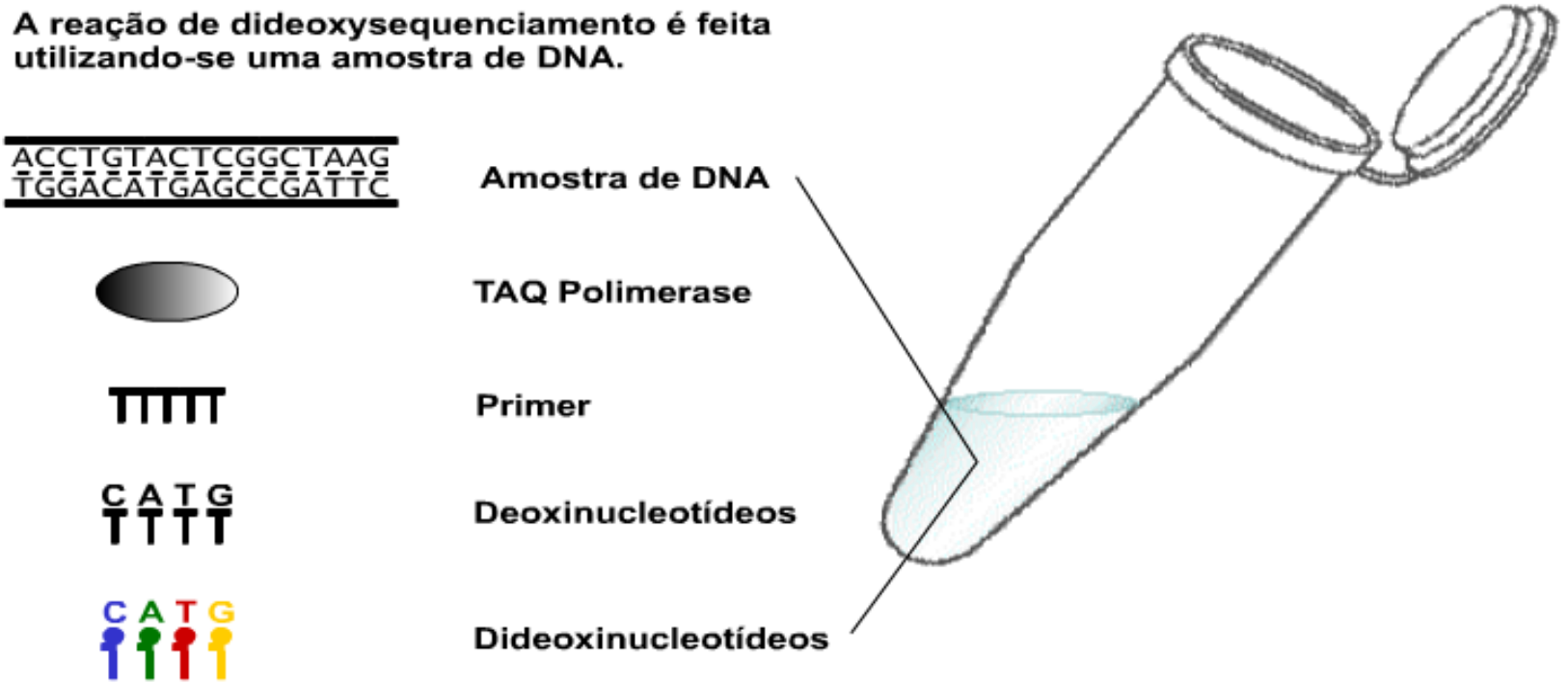
Fonte: A. M. Lesk – Introdução à Bioinformática, Artmed. 2a. Edição, 2008.

REAÇÃO DE SEQUENCIAMENTO

DNA → GEL QUALIDADE → PCR → PURIFICAÇÃO

Ciclo de Sequenciamento

A reação de dideoxysequenciamento é feita utilizando-se uma amostra de DNA.



3’ 5’

GCTAGCCTTGATGGCTAAGTAAGCTGCCA

5’ → 3’

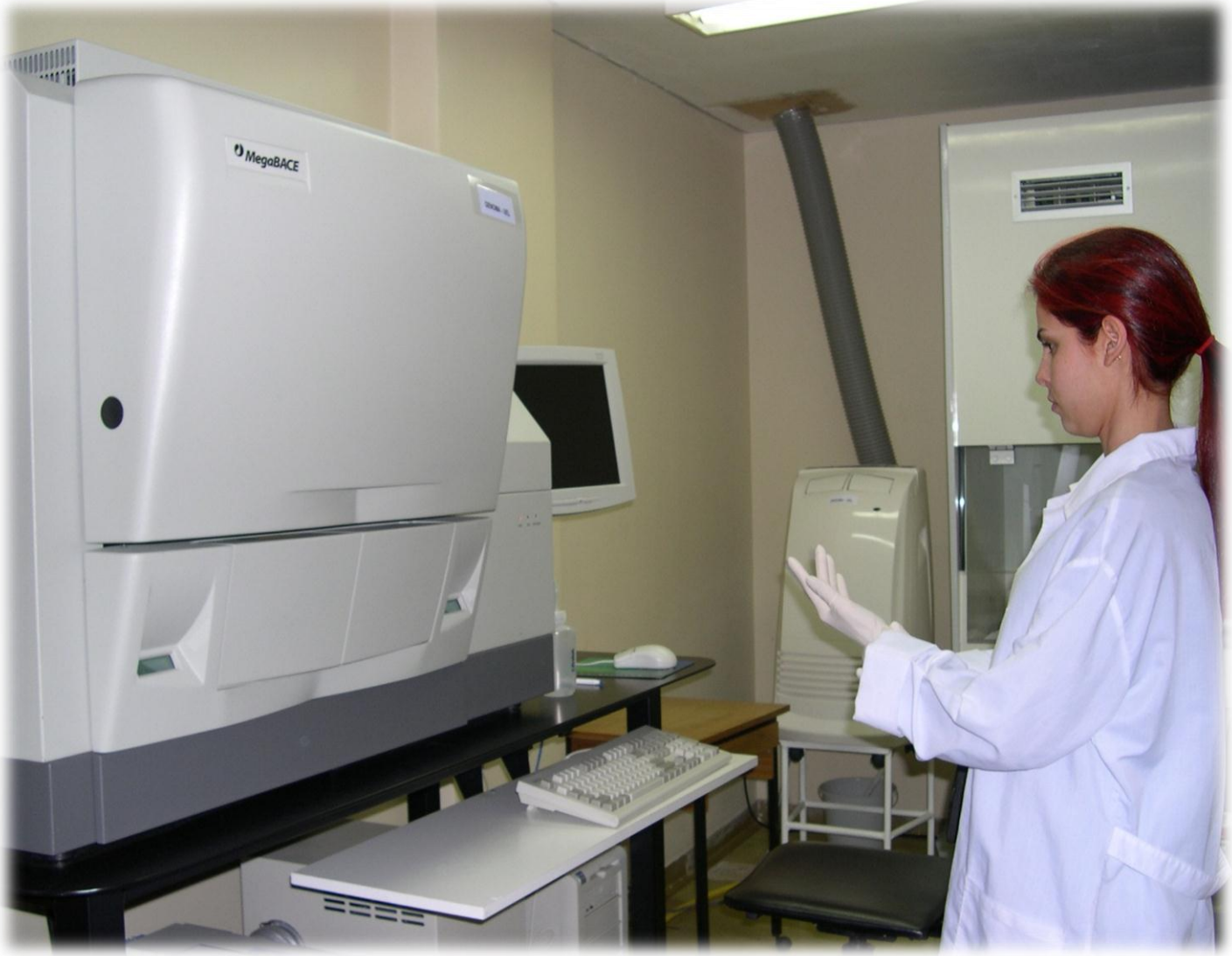
Primer Forward

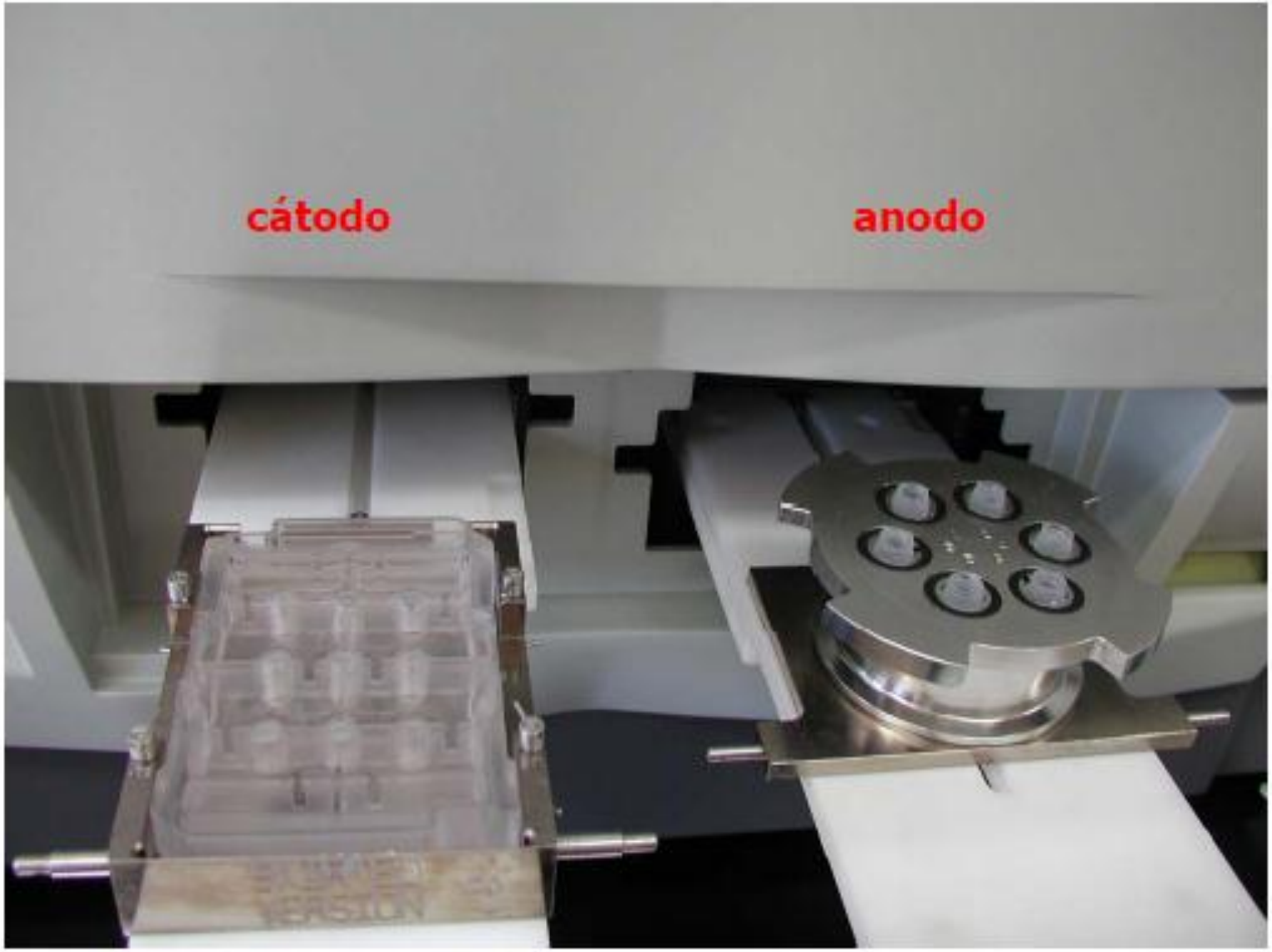
5’ 3’

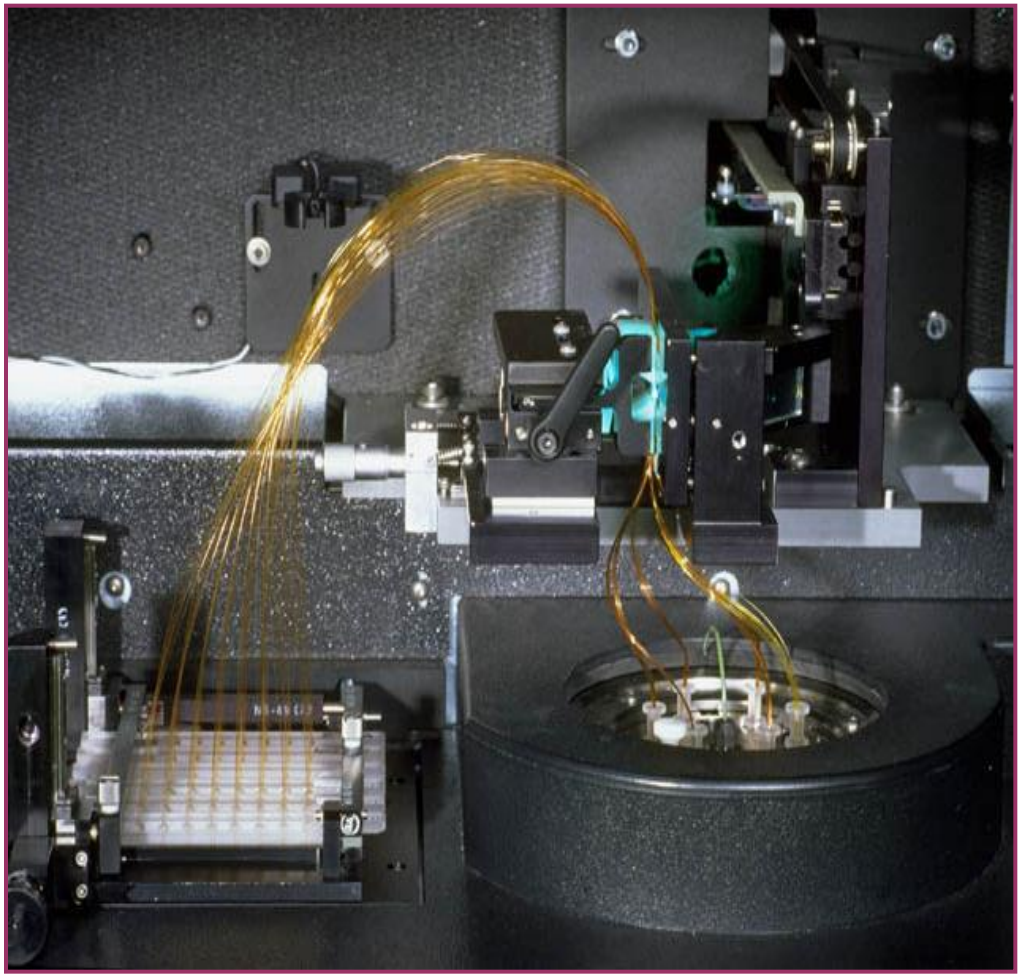
CGATCGGAACTACCGATTCATTCGACGGT

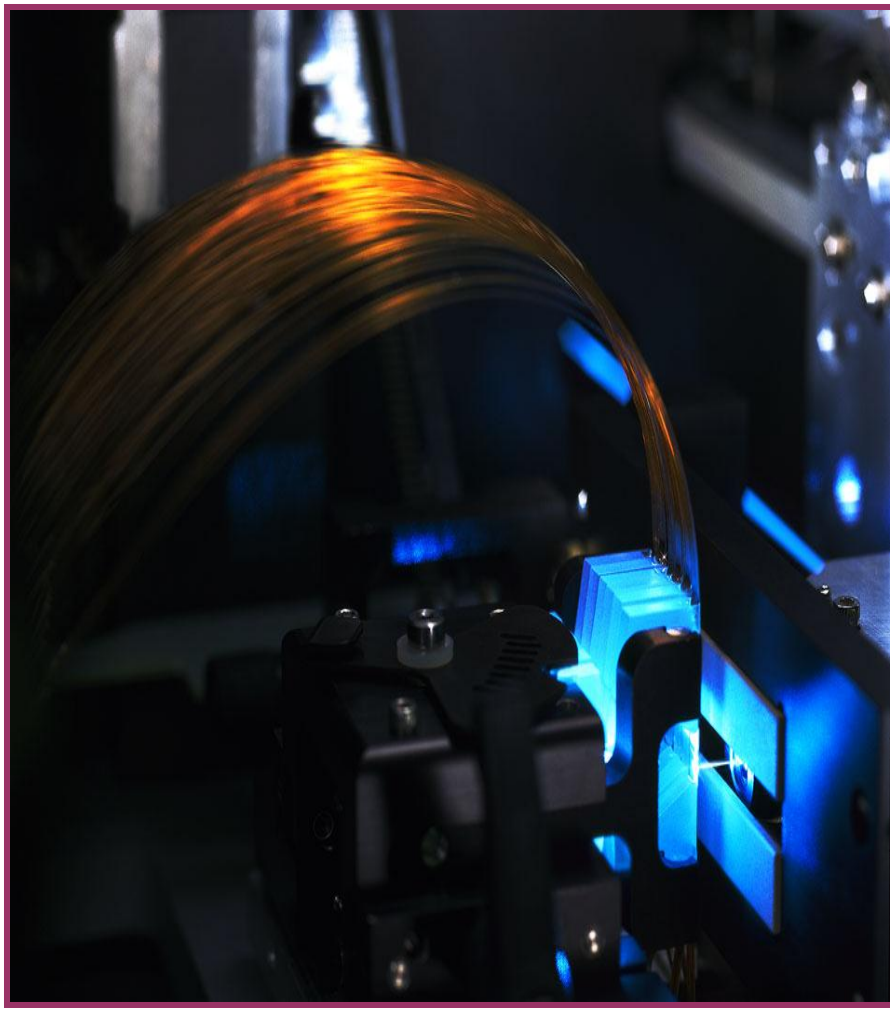
3’ ← 5’

Primer Reverse


MegaBACE


cátodo
anodo

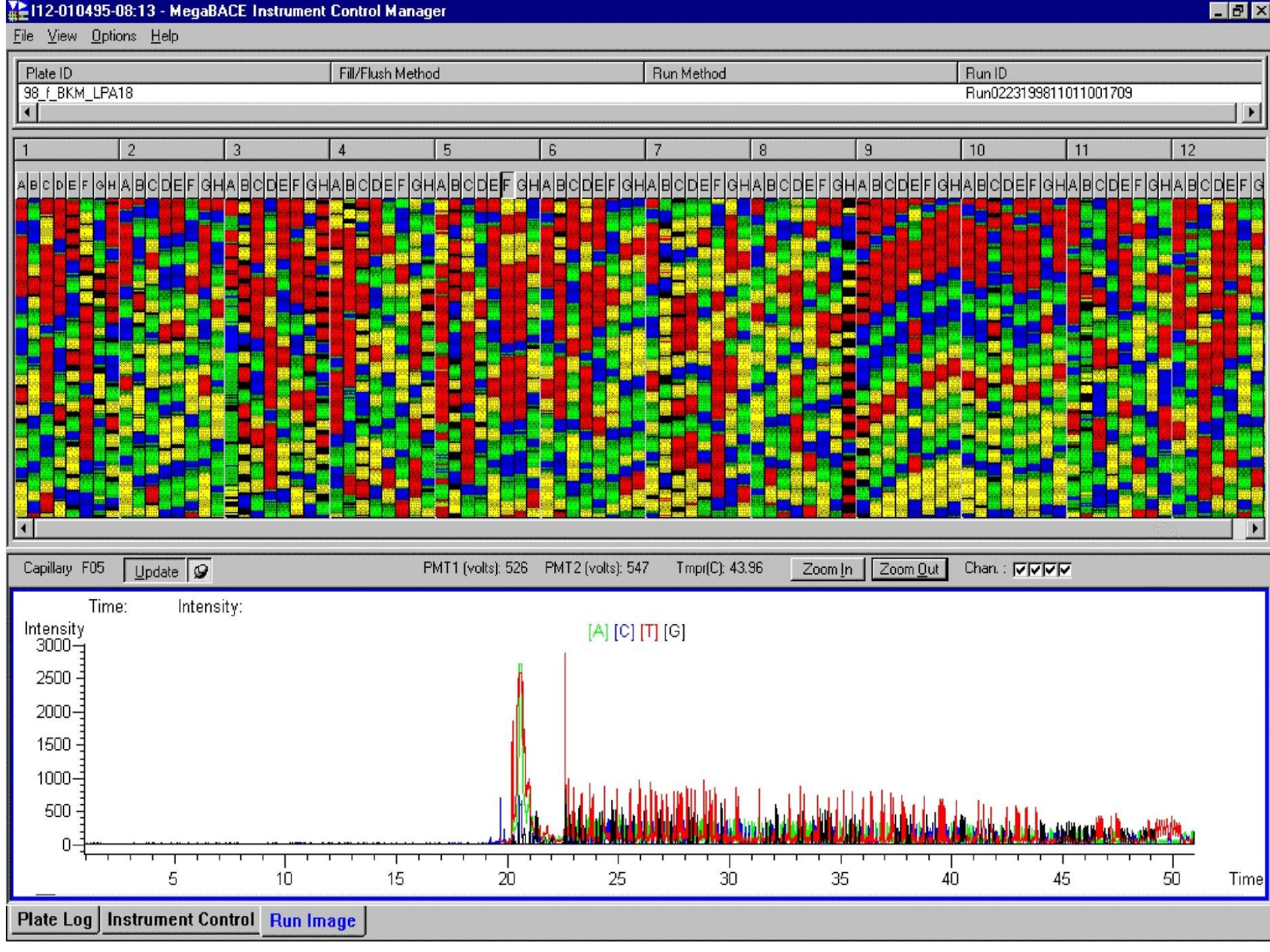




Sala de preparo

Sala de apoio

Sala principal dos sequenciadores


I12-010495-08:13 - MegaBACE Instrument Control Manager
File View Options Help
Plate ID
Fill/Flush Method
Run Method
Run ID
98_f_BKM_LPA18
Run0223199811011001709
1 2 3 4 5 6 7 8 9 10 11 12
Capillary F05
Update
PMT1 (volts): 526
PMT2 (volts): 547
Tmpr(C): 43.96
Zoom In
Zoom Out
Chan. :
Time:
Intensity:
Intensity
[A] [C] [T] [G]
3000
2500
2000
1500
1000
500
0
5 10 15 20 25 30 35 40 45 50
Time
Plate Log
Instrument Control
Run Image

D:\PHRED_CE\Data\DYEnamic_ET_Terminator_Cpl_530_sl_ph_MD1 - MegaBACE ScoreCard

File Edit View Help

molecular dynamics

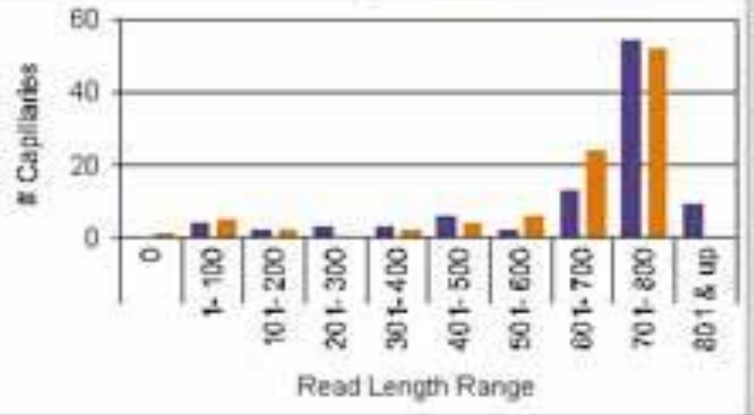
MegaBACE ScoreCard

Plate ID	DYEnamic_ET_Terminator	**Run ID**	Run01	**Run Date**	8/26/99 1:55:49 PM
Run Time (min)	120	**Injection Time (sec)**	30	**Chemistry**	ET Terminators
Run Voltage (kV)	8	**Injection Voltage (kV)**	2	**Base Caller**	Cimarron 1.53 Slim Phredify
Comment					

	1		2		3		4		5		6		7		8		9		10		11		12	
A	767	707	241	319	600	658	250	120	582	671	759	748	4	0	118	5	401	613	788	740	801	705	806	718
B	726	700	680	689	804	742	755	753	702	724	713	687	643	582	250	516	782	720	813	704	682	866	795	726
C	790	686	837	671	752	747	760	688	605	322	703	622	799	778	66	17	772	753	702	649	797	787	747	700
D	742	679	809	569	765	725	763	768	675	679	786	734	793	705	735	724	788	741	744	768	798	749	682	670
E	430	513	639	704	801	749	713	691	24	11	744	747	788	758	652	648	755	713	757	735	777	727	826	747
F	772	552	768	707	384	720	775	752	722	701	323	442	753	707	332	129	749	746	741	689	750	676	753	712
G	791	718	765	707	774	706	774	739	728	704	486	569	784	687	706	685	7	3	820	714	788	719	839	720
H	404	449	671	646	694	722	803	702	517	80	185	452	748	733	473	444	790	645	754	719	782	706	691	731

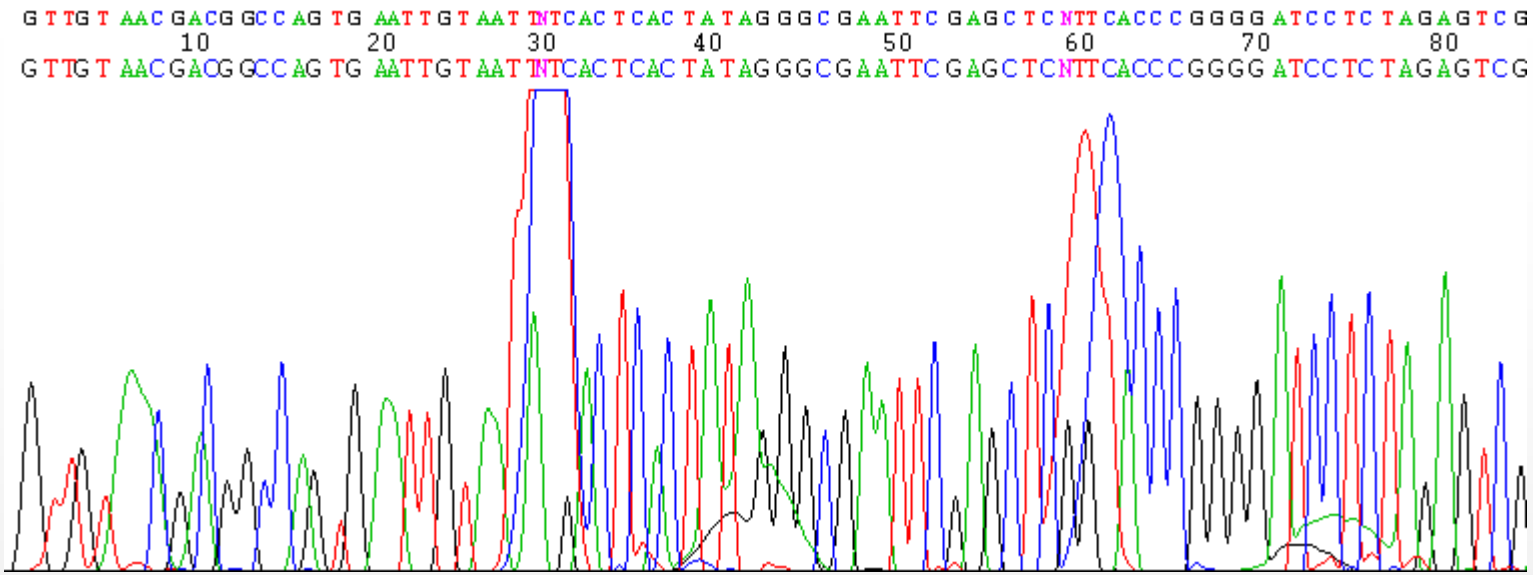
Legend: MD | PHRED (by cutoff, threshold=15) | MD | PHRED

Read Length	MD	PHRED
1	0	1
2-100	4	5
101-200	2	2
201-300	3	0
301-400	3	2
401-500	6	4
501-600	2	6
601-700	13	24
701-800	54	52
801 & up	9	8



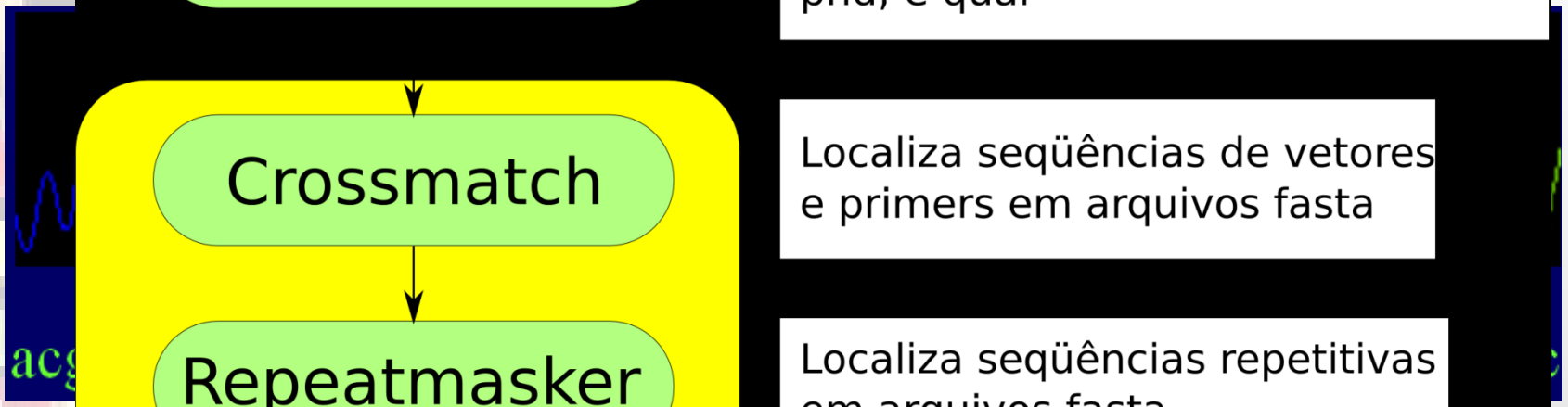
Statistic	Value	Count	%
MD Average	655	96	100
MD Longest Avg	710	87	91
MD > 0 Avg	655	96	100
MD Maximum			839
MD Tot bp/plate			62680
PHRED Average	826	96	100
PHRED Longest Avg	879	89	92
PHRED > 0 Avg	832	95	99
PHRED Maximum			787
PHRED Tot bp/plate			60072

2/25/00 4:54 PM



Receber ➡ Processar ➡ Anotar ➡ Depositar

Para
utiliza
possi
nucle

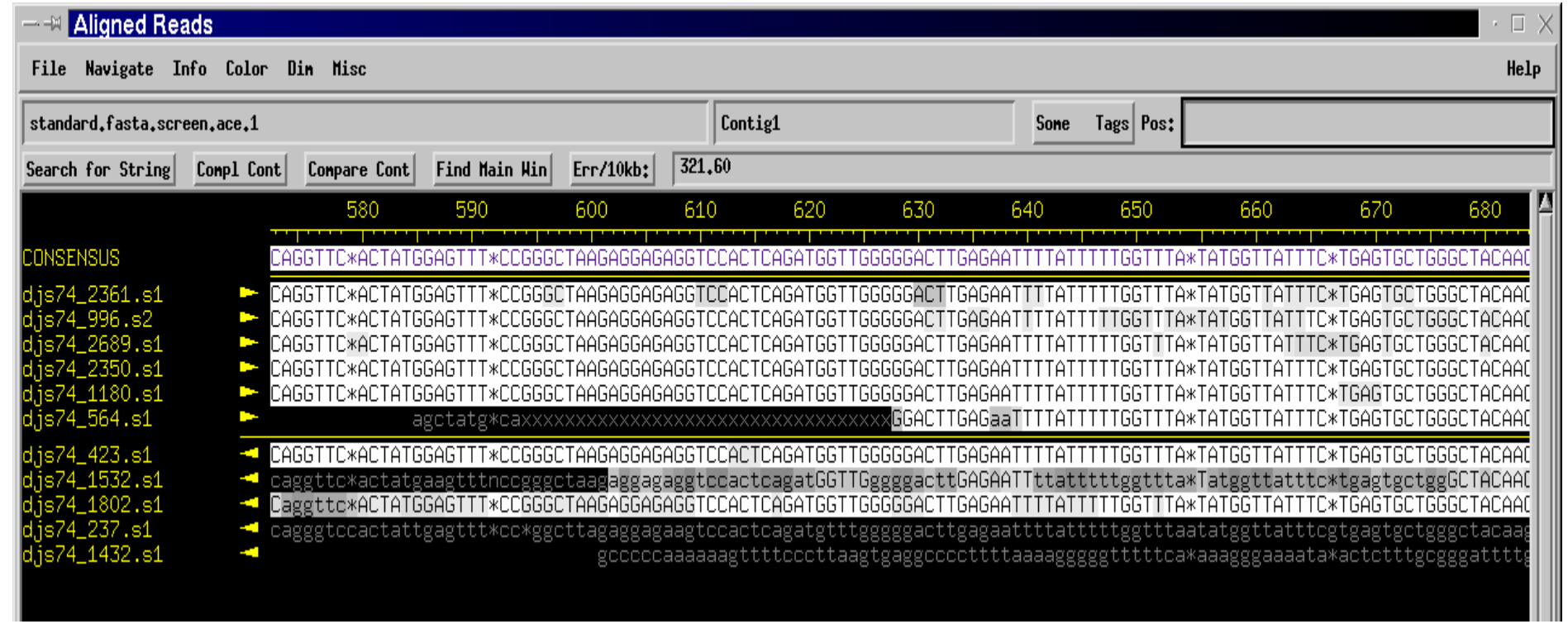


• **PH**
base
as ba

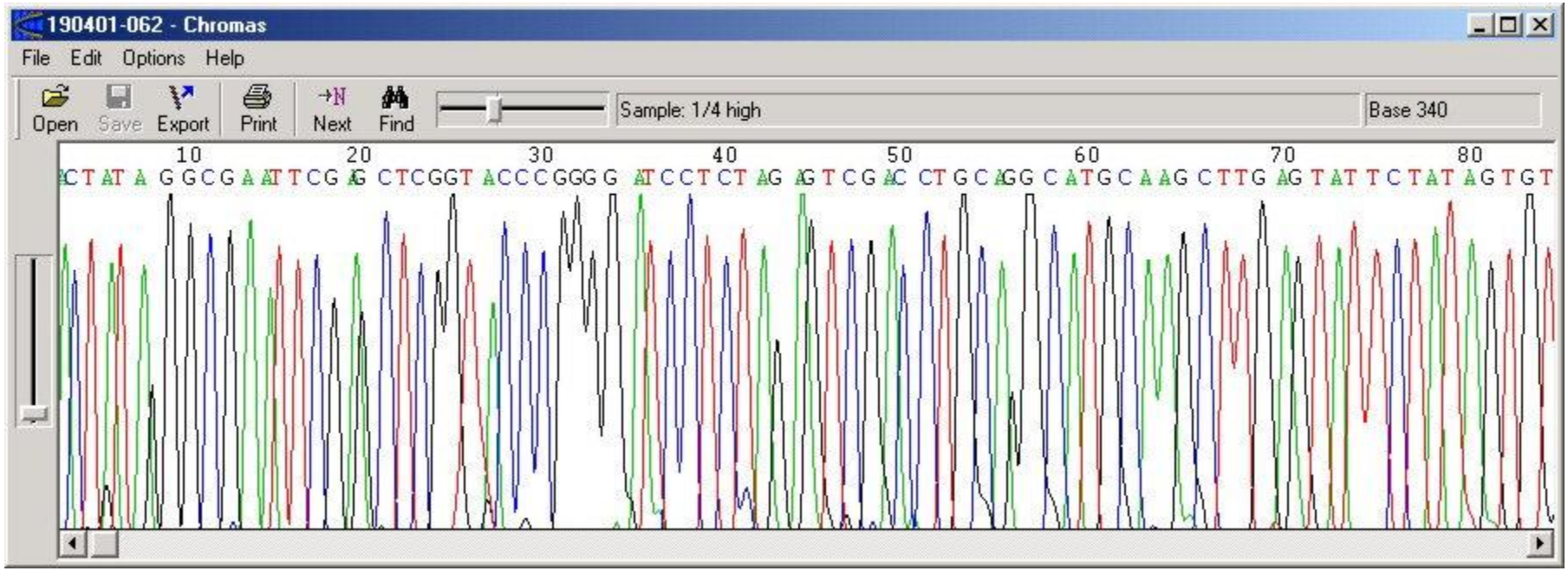
• **PH**
construindo uma sequência contínua ou *contig*.

PHERD – PHRAP - CONSED

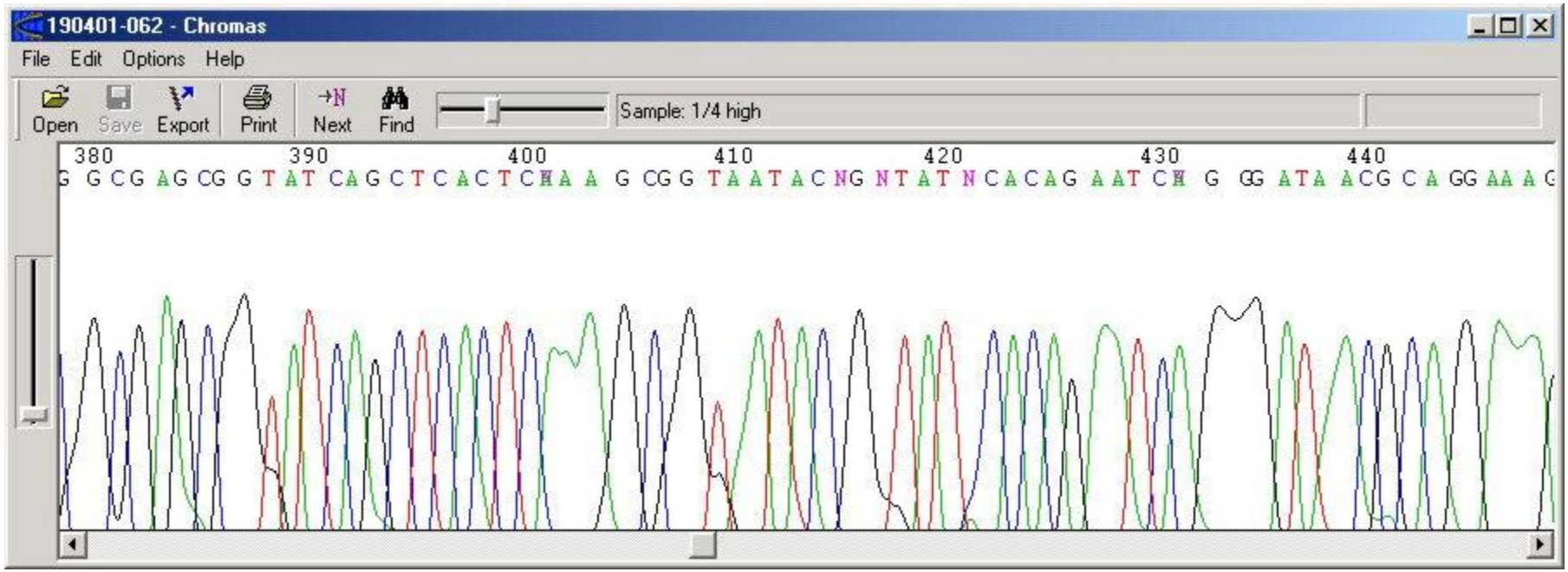
- ✓ Phred = valores de qualidades para as bases
- ✓ Phrap e Cap3 = montagem
- ✓ Consed = visualização



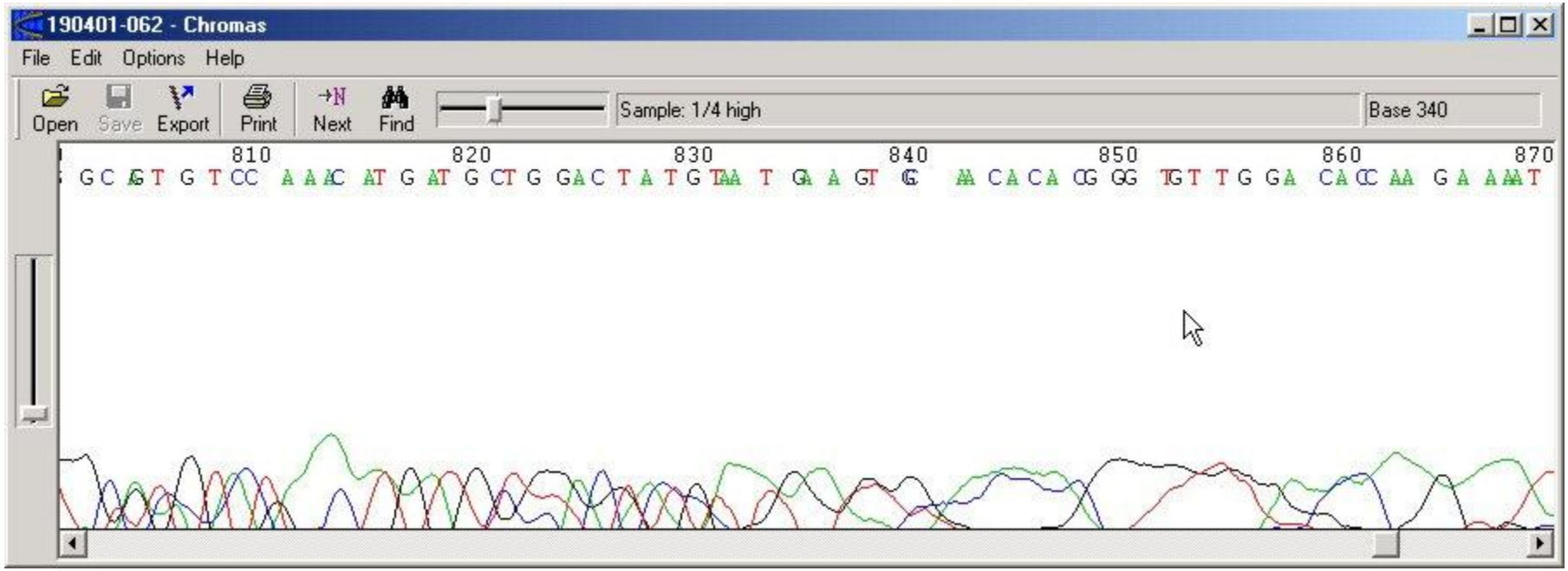
Região de alta qualidade (cromatograma)



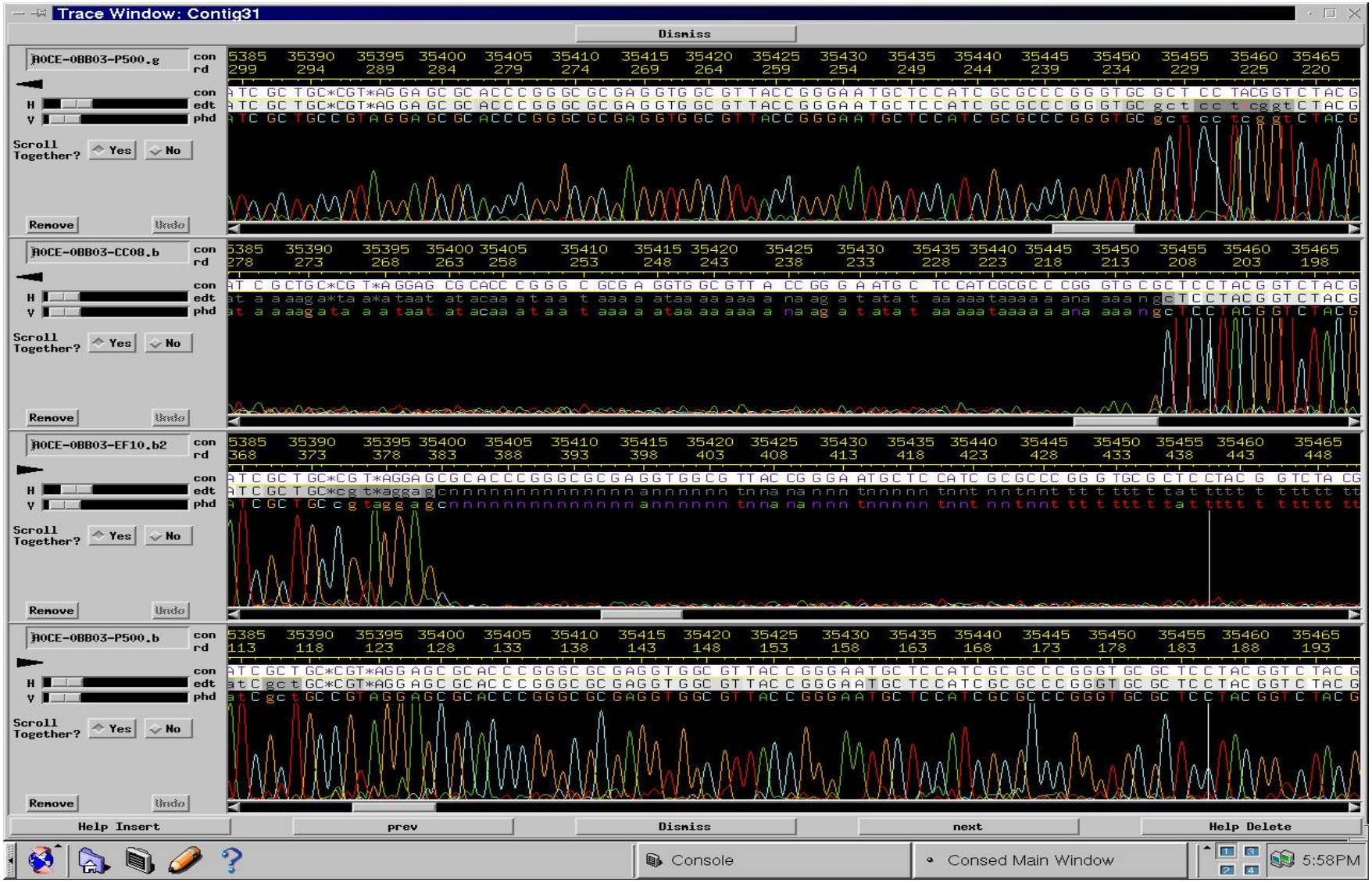
Região de qualidade média

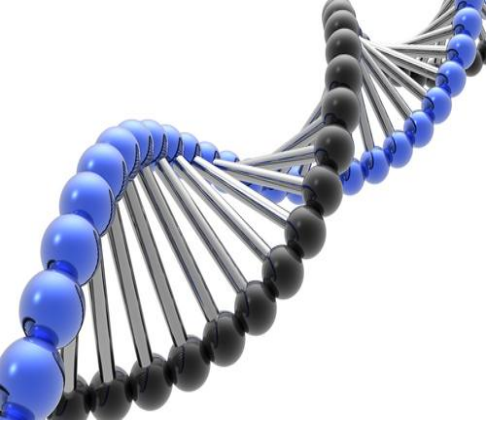


Região de qualidade baixa



“Contaminação”, primers, vetores





ESQUEMA DA FORMAÇAO DE *CONTIGS*

ACGTTGCCTAGTAGATGCTA

reads

GATGCTAACGTTGCCTAGTA

GCCTAGTAGAT

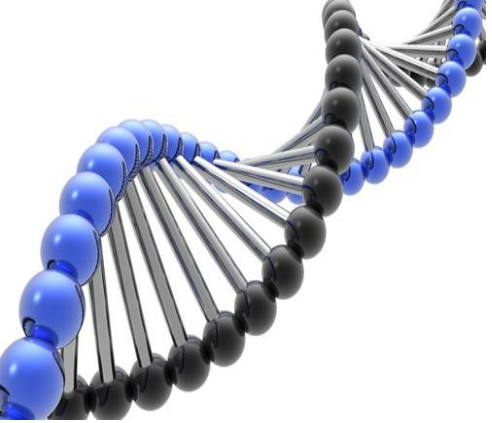
AACGTTGCCTAGTAGCT

clusterização

Contig (contíguo)

ACGTTGCCTAGTAGATGCTAACGTTGCCTAGTAGCT

- Ordenação dos trechos de DNA sequenciados para obtenção da sequencia original
- Inclui verificação de qualidade de bases, marcação de vetores, comparação entre clones e formação de contíguos ate se obter o final



ORFS

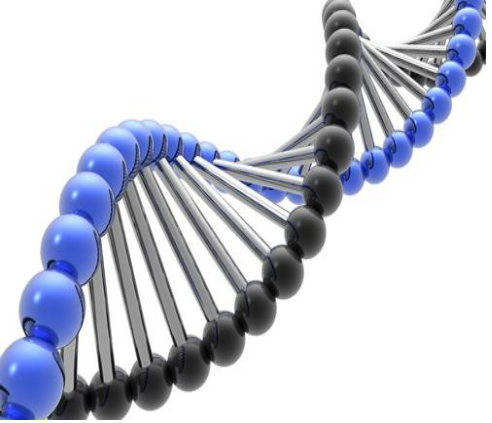
Os ORFs (Open Reading Frames) são genes em potencial.

ATG AAT GCT TGC ACC CCG TCA GGC CTG TAA

ini fim

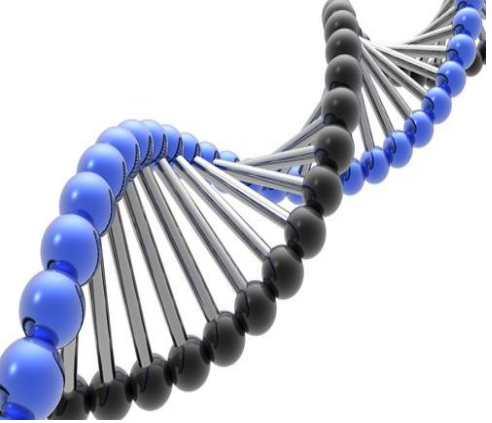
Códon iniciador, região codificadora

e códon terminador.



FASES DE MONTAGEM

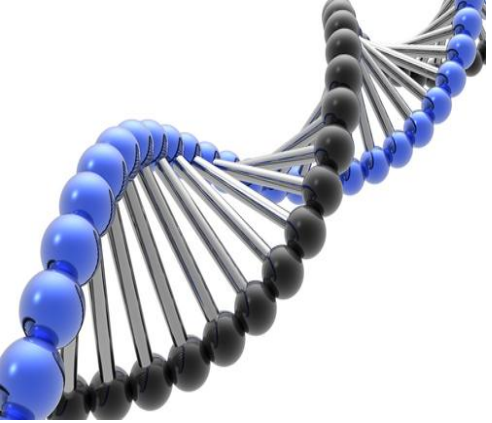
- Remoção das extremidades de baixa qualidade.
- Cálculo de sobreposição de *reads.*
- Remoção das falas sobreposições.
- Construção de *Contigs.*
- Alinhamento múltiplo e geração de consenso.



Anotação de genes

Anotar é postular função ao produto de um ORF.

Utilizam-se diversos programas de comparação com dados genéticos conhecidos.



ANOTAÇÃO GENÔMICA

- É um processo que envolve: Localização de ORF’s

  Identificação de genes

  Pesquisa de estrutura gênica

  Atribuição de função biológica

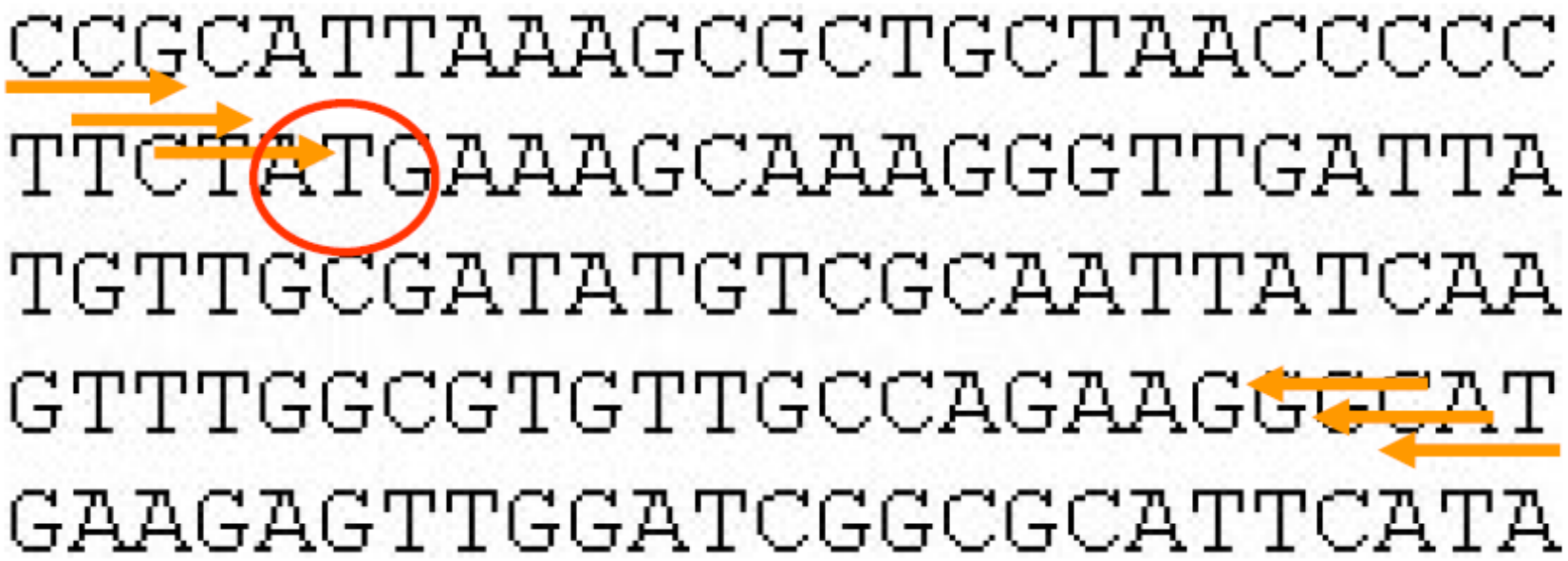
ANOTADOR

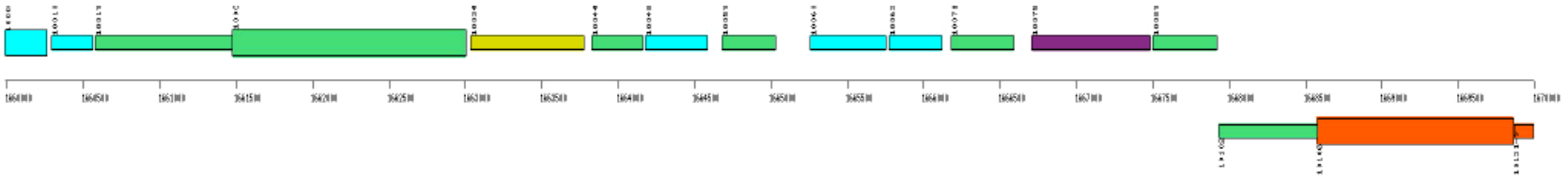
ANOTAÇÃO DE NUCLEOTÍDEOS → **ANOTAÇÃO DE PROTEÍNAS** → **ANOTAÇÃO DE PROCESSOS**

Onde estão os genes? | Quais são os genes? | Cómo os genes interagem?

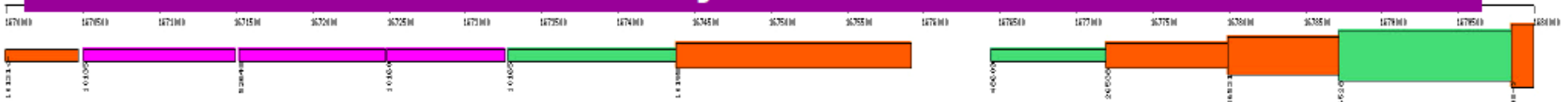
GLIMMER - GENEMARK

Predições de ORF – Open Reading Frame "6 frames"

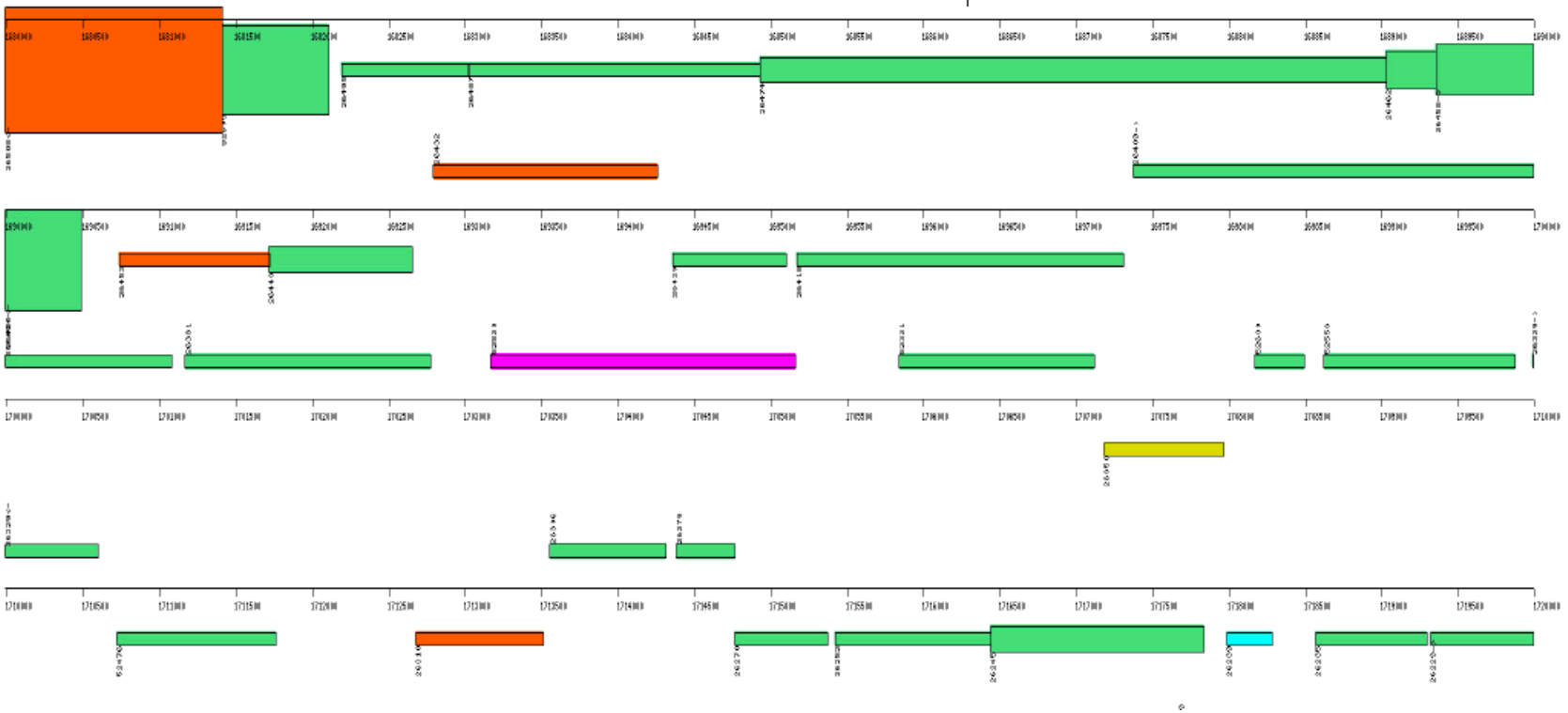




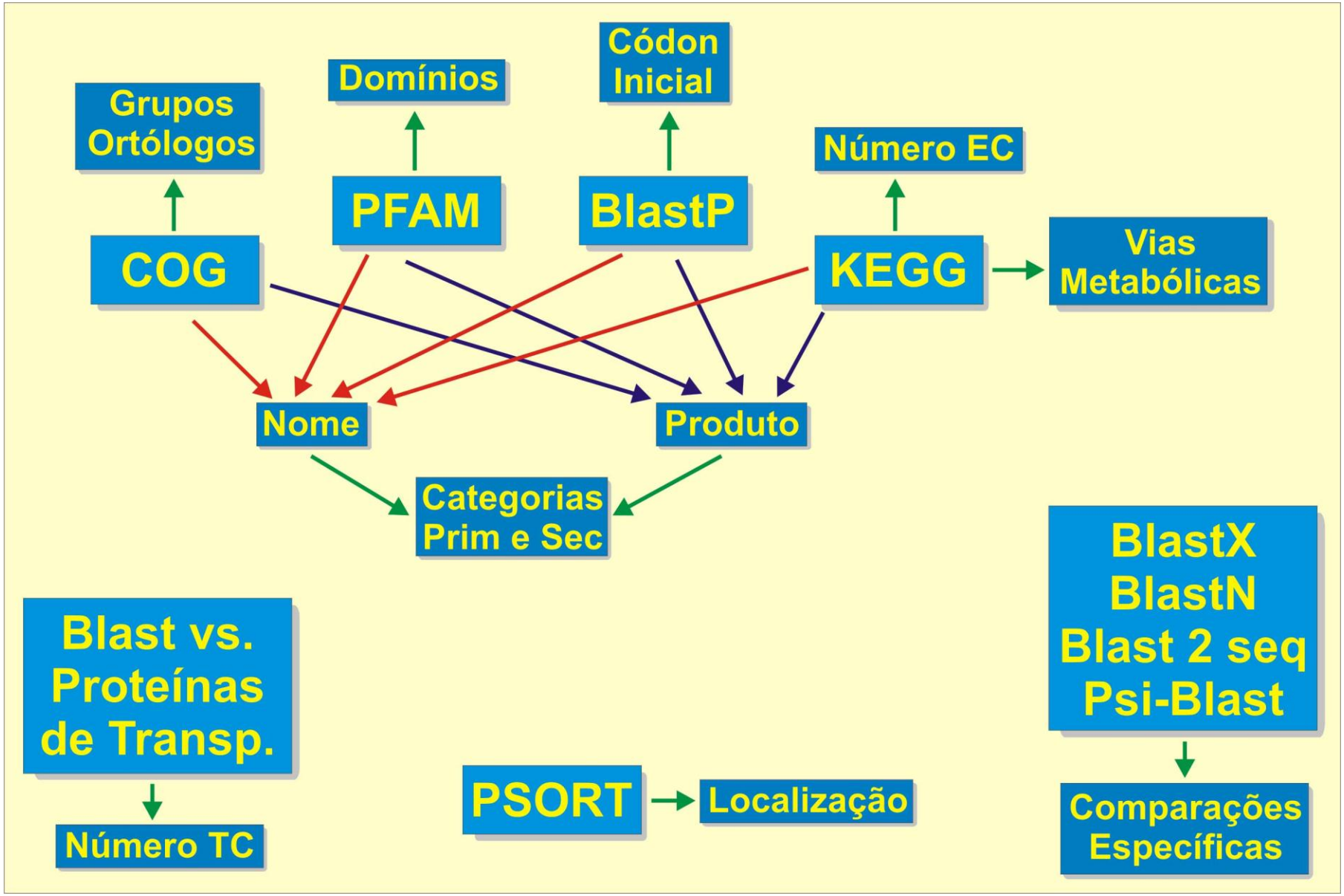
CONTIG APÓS A ANOTAÇÃO MANUAL .....



O anotador decide qual ORF tem que eliminar ou validar. Requere-se de conhecimento e de muito critério.

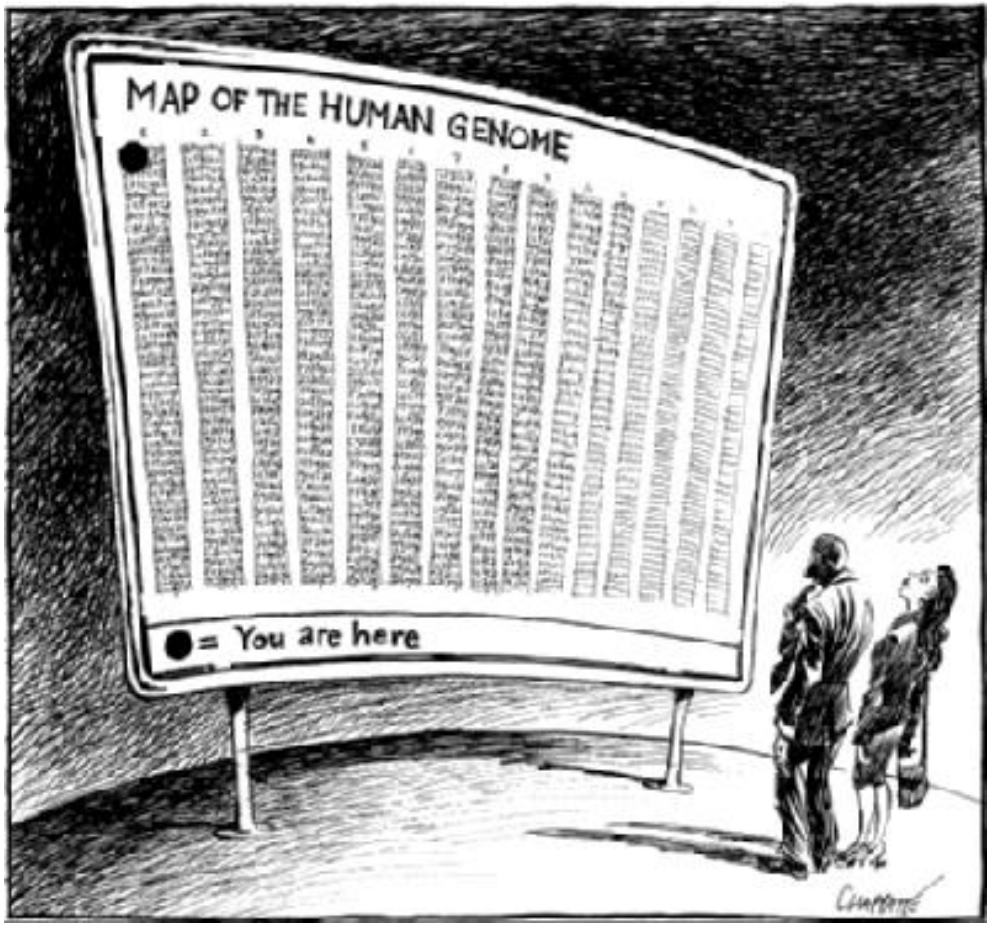


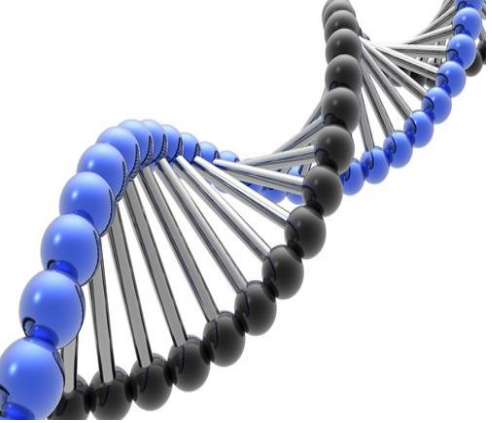
Anotação – ferramentas



PORÇ

E agora ?



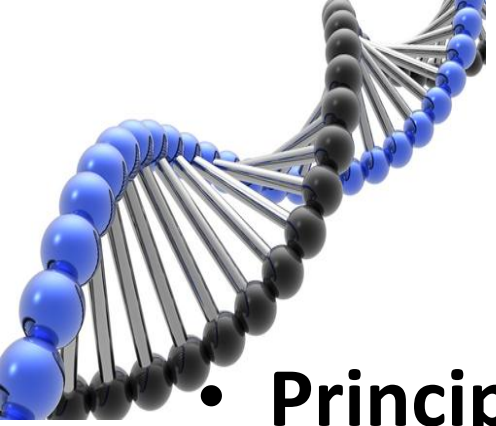


NASCIMENTO DA BIOINFORMÁTICA

É o estudo da estrutura inerente à informação e aos sistemas biológicos. Une dados biológicos (genomas) à teoria analítica e às ferramentas práticas da matemática e da ciência de computação.

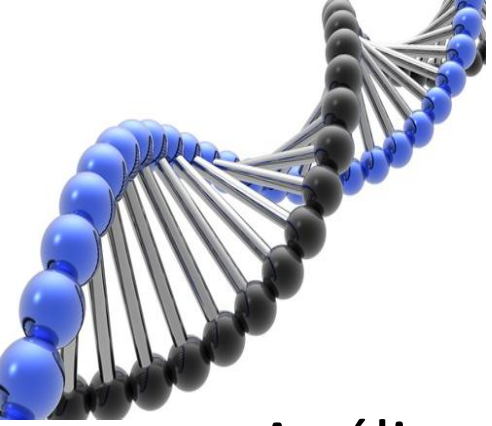
- Advento do Projeto Genoma Humano.
- Sequenciadores automáticos, levou a explosão de informações.
- Há treze anos, o termo nem sequer existia.

Bases da Bioinformática



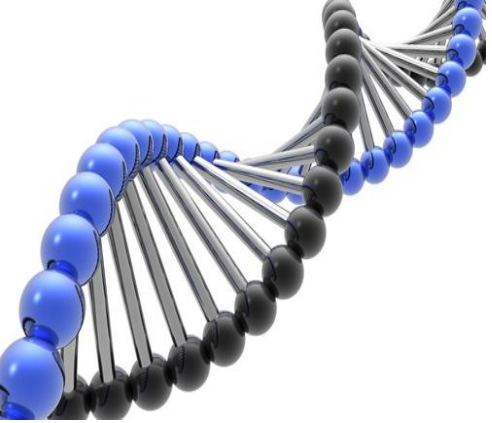
Bioinformática “Clássica”

- **Principais interesses** :
  - Armazenar
  - Recuperar
  - Analisar
  - Predizer
  - Simular

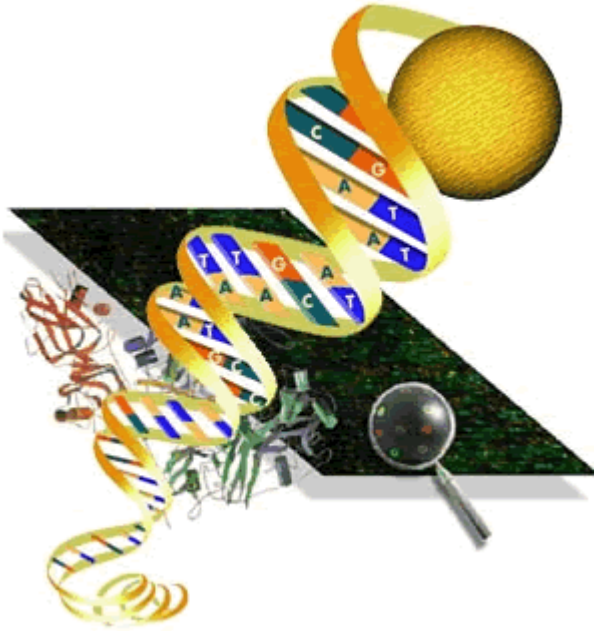


Bioinformática Atual

- Análise dos resultados obtidos através da Genômica.
- Análise dos dados obtidos através de novas técnicas de laboratório.
- Desenvolvimento de modelos de simulação de processos biológicos.
- Desenvolvimento de metodologias para o reconhecimento de padrões que determinam um determinado fenótipo.
- Análise entre os dados clínicos de pacientes e os obtidos através da pesquisa genômica e proteômica.



Centros de Pesquisa



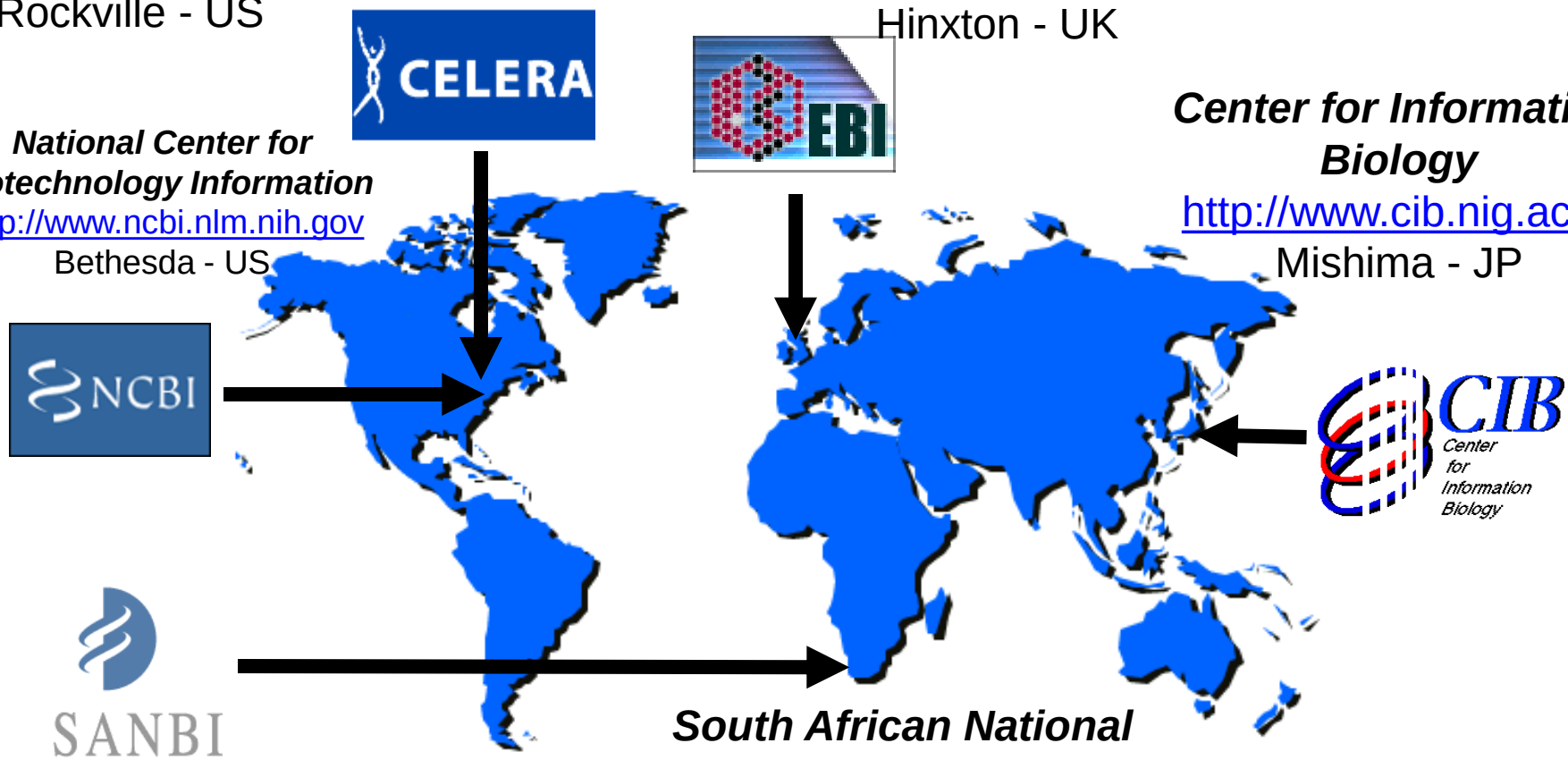
No mundo...

Celera Genomics Group
http://www.celera.com
Rockville - US

European Bioinformatics Institute
http://www.ebi.ac.uk
Hinxton - UK

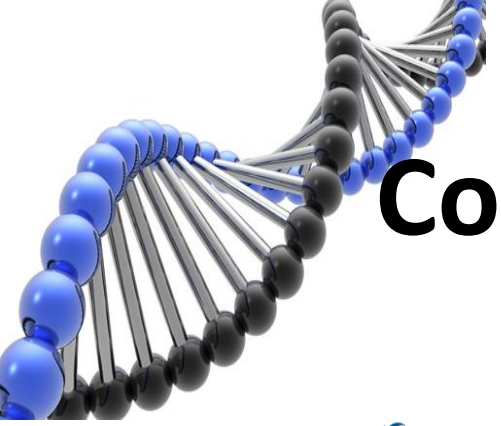
National Center for Biotechnology Information
http://www.ncbi.nlm.nih.gov
Bethesda - US

Center for Information Biology
http://www.cib.nig.ac.jp
Mishima - JP



South African National Bioinformatics Institute
http://www.sanbi.ac.za
Tygerberg - ZA

Companhias de Bioinformática



Abbott Laboratories
Agilent Technologies
AstraZeneca INTERNATIONAL
Base4 BIOINFORMATICS
Biosentients
Computercraft
EMORY UNIVERSITY
GENE DATA
accelrys
AMGEN Click For Career Opportunities
AVAKI
Bayer Pharmaceutical Division
Boehringer Ingelheim
CELERA Genomics
CuraGen Corporation
EXELIXIS pharmaceuticals, inc.
GENENCOR INTERNATIONAL
adolor CORPORATION
Ardais
Aventis
BioDiscovery, Inc.
Bristol-Myers Squibb Company
CIPHERGEN
DU PONT MERCK
GENAISSANCE PHARMACEUTICALS
Genentech, Inc.
GENSET
Genomics Institute of the Novartis Research Foundation
immunex
LEXICON
MERCK
NIH Research & Training Opportunities
Ocimum Biosolutions Ltd.
Pfizer
Strand Genomics
Virtual Genetics
genzyme
GlaxoSmithKline
Incyte Pharmaceuticals
Lilly
Monsanto
NOVARTIS
PARACEL
Roche
Syn•mics
Wyeth
GEOSPIZA INC. Software for Molecular Biology
ILS
Incellico Analytic software for the life sciences
bioscience LION AG
neomorphic
Novo Nordisk
Partek
Schering-Plough Research Institute
3rd Millennium

http://www.bioplanet.com/bioinformatics_companies.htm

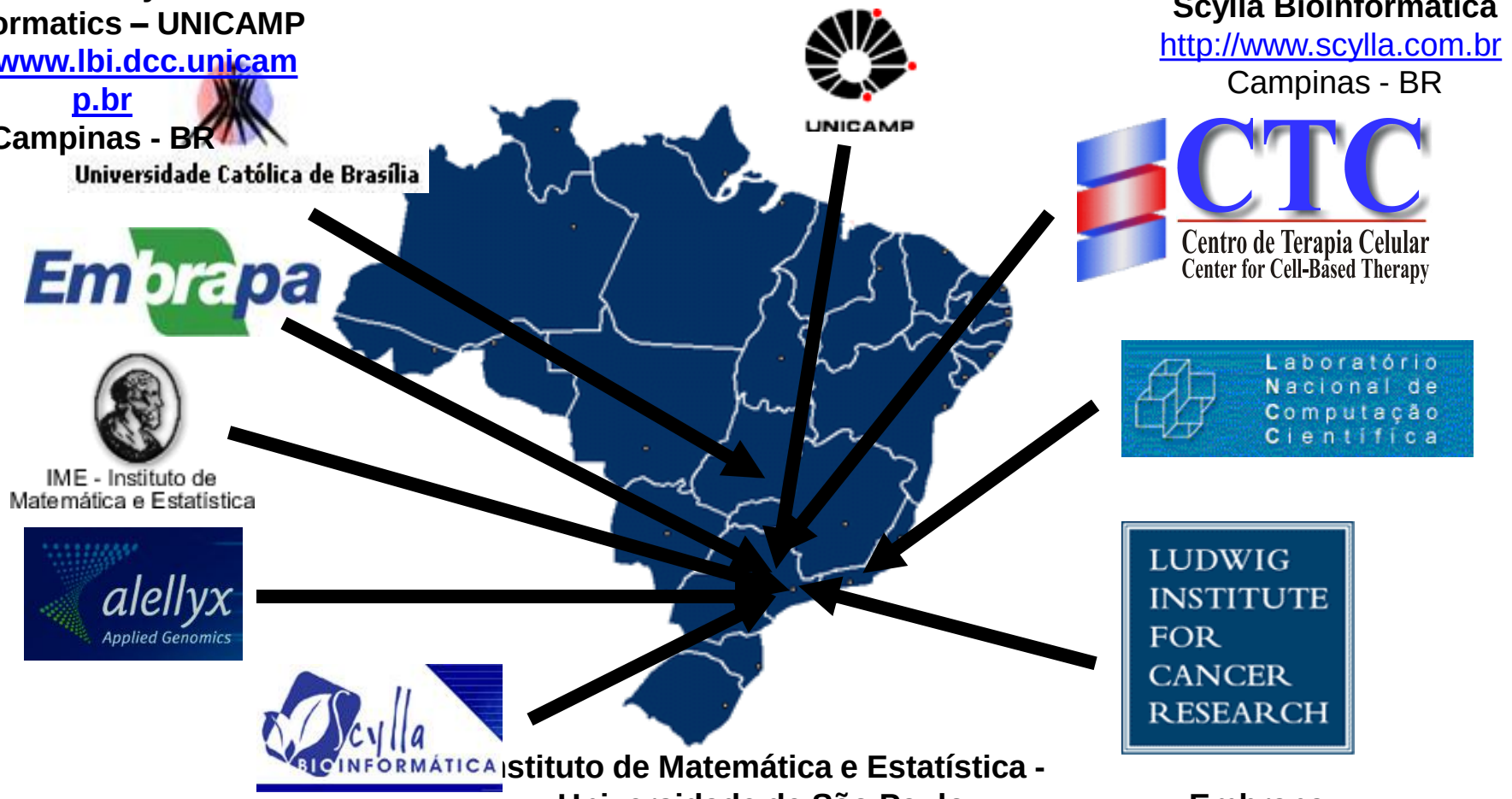
Laboratório Nacional de Computação Científica
http://www.lncc.br
Petrópolis - BR

***Bioinformatics Laboratory* - Universidade Católica de Brasília**
http://bioinformatica.ucb.br
Brasília - BR

Centro de Terapia Celular (BiT)
http://ctc.fmrp.usp.br
Ribeirão Preto - BR

Laboratory for Bioinformatics – UNICAMP
http://www.lbi.dcc.unicamp.br
Campinas - BR

Scylla Bioinformática
http://www.scylla.com.br
Campinas - BR



Instituto de Matemática e Estatística - Universidade de São Paulo
http://www.bioinfo.usp.br
São Paulo - BR

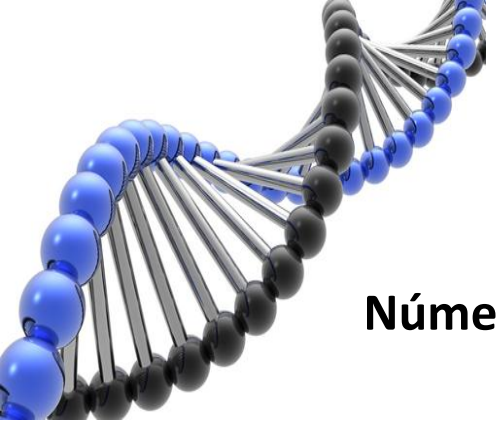
Embrapa
http://www.nbi.cnptia.embrapa.br
Campinas - BR

Ludwig Institute for Cancer Research – São Paulo Branch
http://www.ludwig.org.br
São Paulo - BR

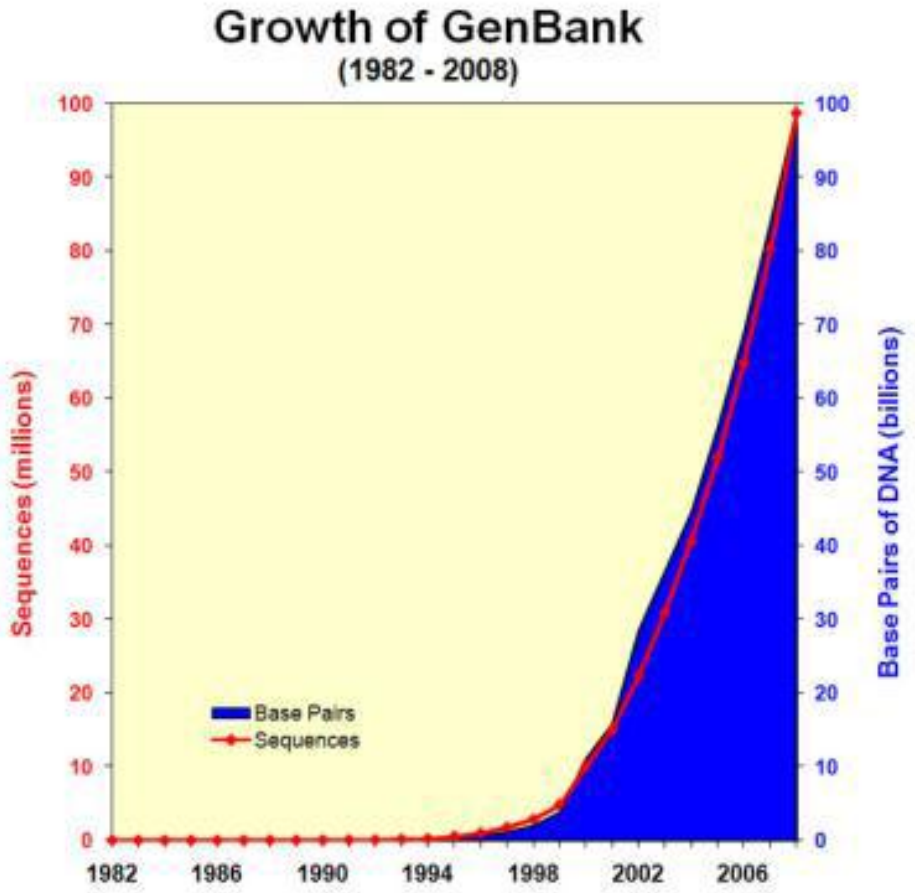
Alellyx Applied Genomics
http://www.alellyx.com.br
Campinas - BR

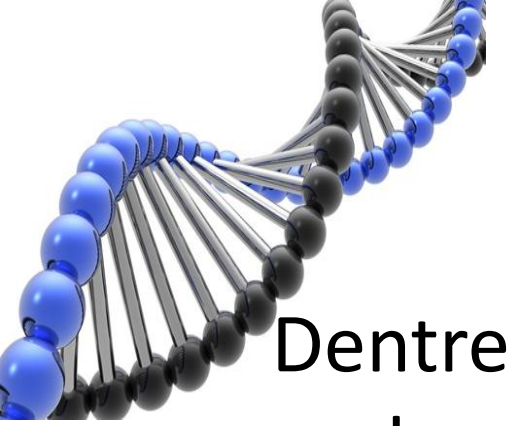
Tipos de dados:

- Bancos de dados de DNA genômico
- Bancos de dados de cDNAs: bibliotecas de cDNAs parciais e completos
- Bancos de dados de ESTs - *Expressed Sequence Tags*
  (cDNAs sequenciados só de um lado)
- Bancos de dados de STSs (Sequence-tagged sites): dbSTS – útil para estudos de mapeamento
- Sequências Genômicas em Larga Escala (HTGS): dados de projetos de sequenciamento ainda não concluídos
- Bancos de Dados de Proteínas: sequências de proteínas de várias fontes SwissProt, PIR, PRF, PDB, e traduções de regiões codificadoras do GenBank e RefSeq



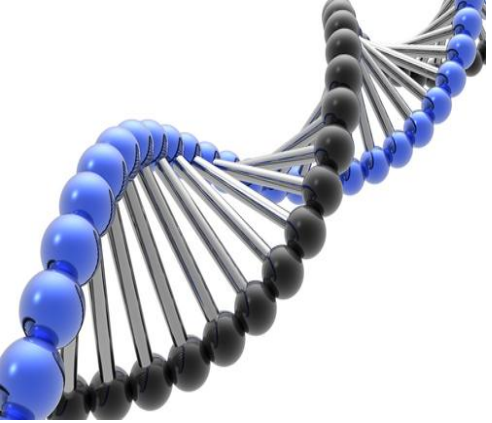
Número de sequências depositadas no GenBank (15/02/2012)





Dentre as características da Bioinformática, pode-se citar:

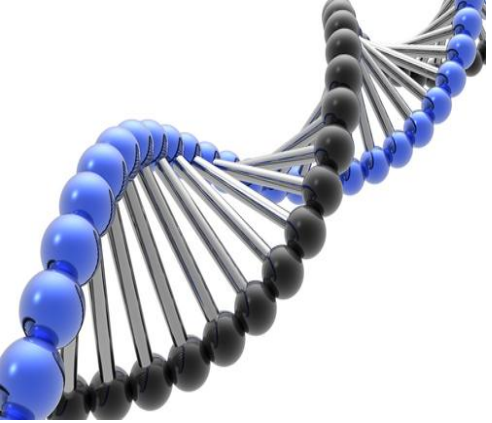
- O recebimento das sequências
- O tratamento de sequências e a montagem do genoma
- A anotação do genoma.
- Base para novas hipóteses



Importância

- Genoma Humano: previsto para ser desenvolvido e concluído em 15 anos, foi antecipado, em cerca de 5 anos.

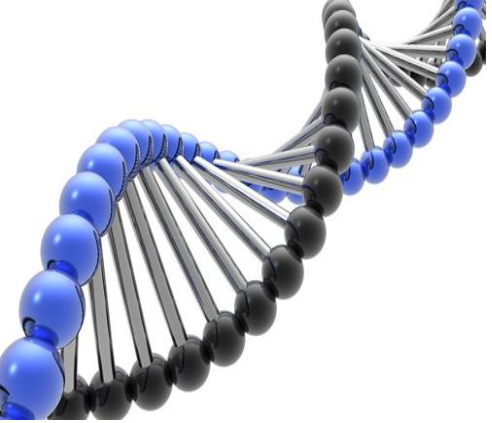
- Hoje, um novo gene, com 12 mil bases tem sua sequência decifrada em 1 minuto, há 3 anos atrás a mesma tarefa levaria 20 minutos.



.... BIOINFORMÁTICA

União de diversas linhas de conhecimentos: engenharia de softwares, matemática, estatística, ciências da computação e a biologia molecular.

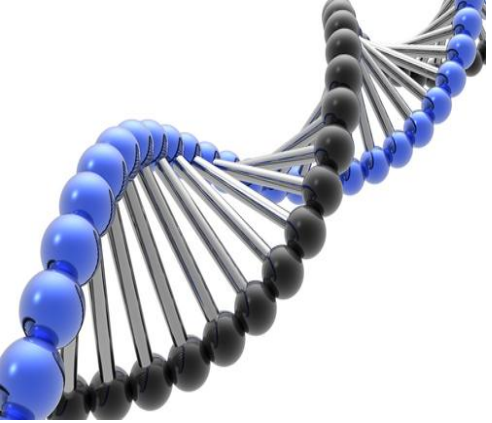
- **Estudos...**
- **Mapeamento;**
- **Evolução;**
- **Comparação de sequências;**
- **Modelagem gênica;**
- **Sequência DNA-> estrutura;**
- **Modelagem molecular;**
- **Comparação de estruturas;**
- **Sequências/estruturas->função;**
- **Expressão gênica e redes genéticas/metabólicas;**
- **Banco de dados**
- **Visualização e interação.**



Sistemas Operacionais

- Windows
- Unix (mais usado)
- MacOs

- **Mais confiáveis**
- **Gerenciam melhor grande quantidade de dados**
- **em código aberto**

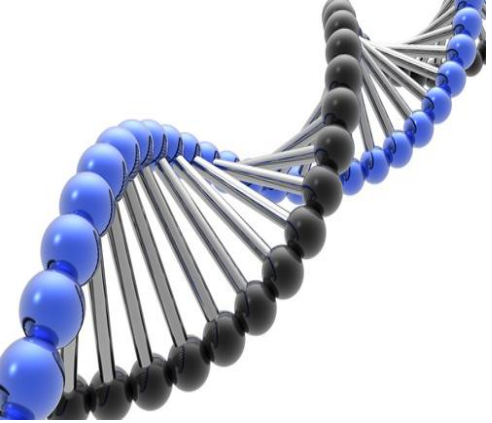


Linguagens de Programação

- Basic
- Cobol
- Perl
- Dbase
- Java
- Fortran

PERL

- **Linguagem de alto nível**
- ***Freeware***
- **Módulos:**
  - ✓**Bioperl**
  - ✓**Biographics**
- **Interconectividade c/ bancos de dados**
  - ✓**Perl-DBI**

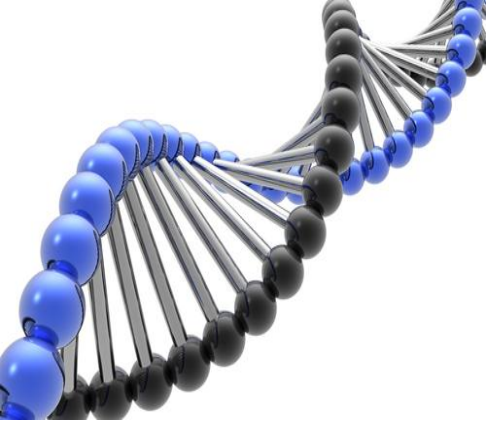


BANCOS GENÔMICOS

Qual a função desses Bancos Genômicos?

- Armazenar todas as sequências geradas pelos genomas (nucleotídeos, proteínas ...).

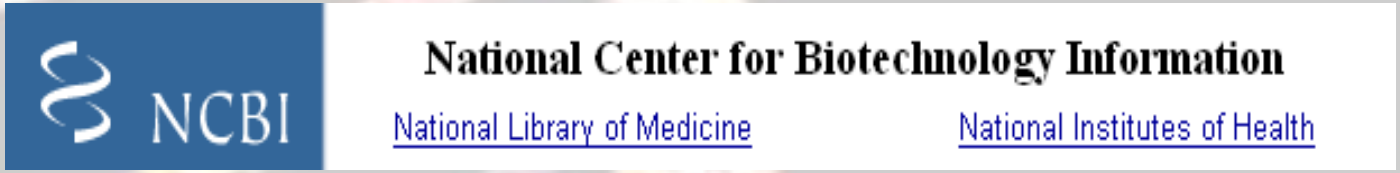
- Organizar de maneira acessível para que não ocorra redundância nas pesquisas científicas.

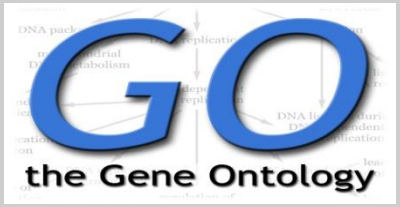


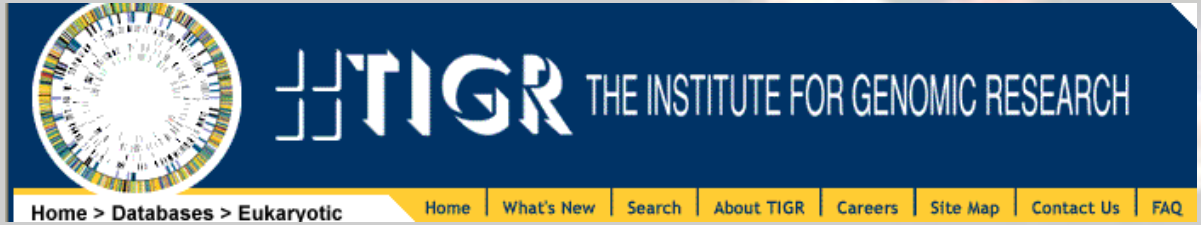
Razões para se usar um banco de sequências

- **Eu acabei de obter uma sequência. O que é sabido à respeito desta sequência? Ela é única?**
- **Eu tenho uma sequência única. Ela tem similaridade com alguma outra sequência de função conhecida?**
- **Eu encontrei uma nova proteína em um determinado organismo. Existe um ortólogo conhecido?**
- **Eu decidi trabalhar com um gene novo. Eu não tenho como obter um clone contendo a sequência deste gene. Eu preciso da sequência do cDNA para fazer uma PCR.**

Bancos Genômicos de acesso público



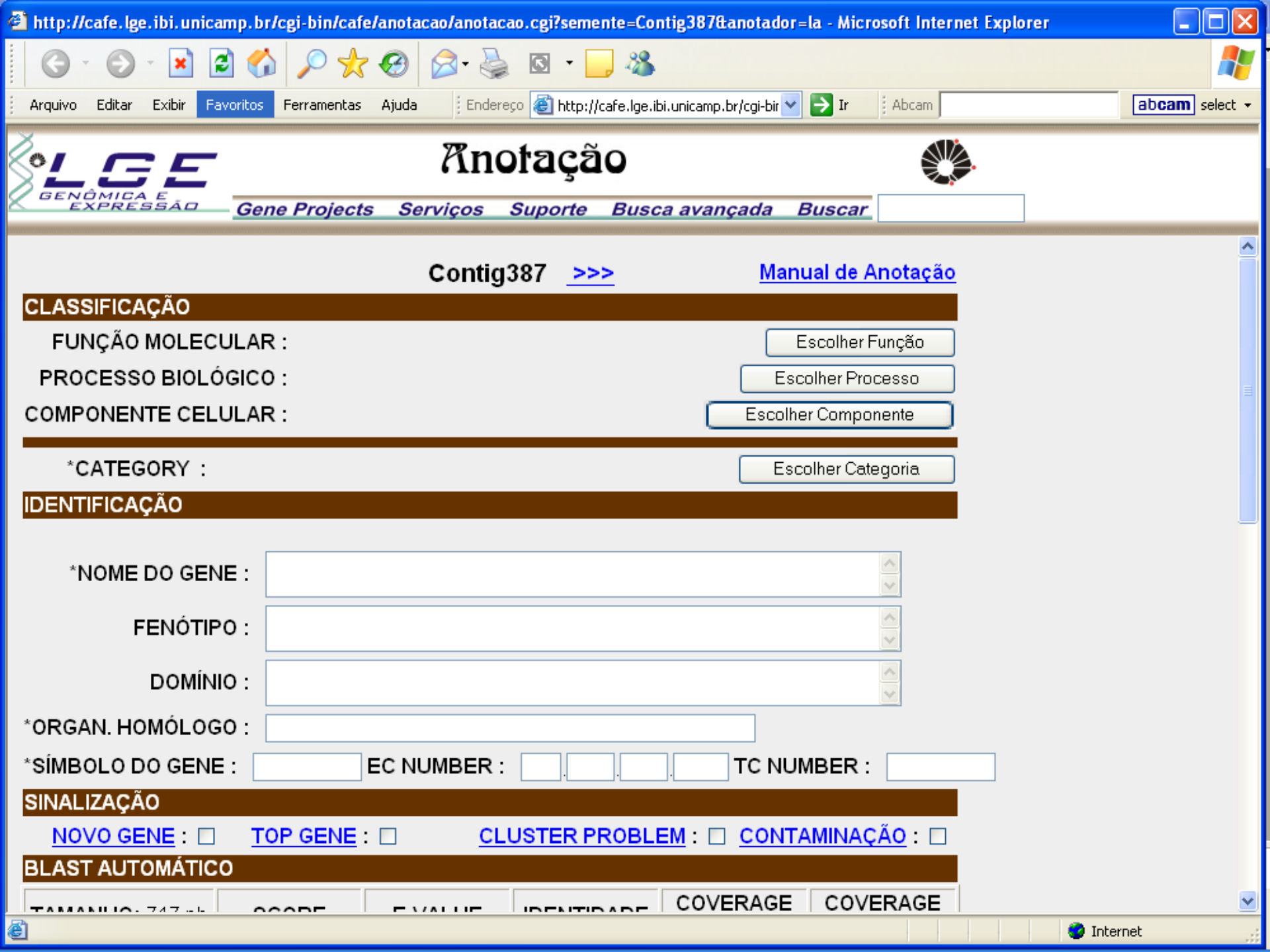




1000 espécies de referência, já completou o sequenciamento e montagem do genoma de 95 espécies, sendo que outras 505 estão com seus genomas em fase final de sequenciamento e/ou montagem.

Reconhecida como talvez a maior plataforma de genômica e bioinformática do mundo, com capacidade de sequenciar até 30.000 genomas humanos por ano, cientistas do *BGI* quebraram vários paradigmas desenvolvendo tecnologias que hoje em uso eram tidas como impossíveis de funcionar.

Bioinformática - LGE


http://cafe.lge.ibi.unicamp.br/cgi-bin/cafe/anotacao/anotacao.cgi?semente=Contig387&anotador=la - Microsoft Internet Explorer
Arquivo Editar Exibir Favoritos Ferramentas Ajuda
Endereço http://cafe.lge.ibi.unicamp.br/cgi-bir
Ir
Abcam
abcam select
LGE GENÔMICA E EXPRESSÃO
Anotação
Gene Projects Serviços Suporte Busca avançada Buscar
Contig387 >>>
Manual de Anotação
CLASSIFICAÇÃO
FUNÇÃO MOLECULAR :
Escolher Função
PROCESSO BIOLÓGICO :
Escolher Processo
COMPONENTE CELULAR :
Escolher Componente
*CATEGORY :
Escolher Categoria
IDENTIFICAÇÃO
*NOME DO GENE :
FENÓTIPO :
DOMÍNIO :
*ORGAN. HOMÓLOGO :
*SÍMBOLO DO GENE :
EC NUMBER :
TC NUMBER :
SINALIZAÇÃO
NOVO GENE :
TOP GENE :
CLUSTER PROBLEM :
CONTAMINAÇÃO :
BLAST AUTOMÁTICO
COVERAGE
COVERAGE
Internet

GENOPAR - Contig Annotation Page [MF lab] - Microsoft Internet Explorer

Arquivo Editar Exibir Favoritos Ferramentas Ajuda

Voltar Avançar Parar Atualizar Página inicial Pesquisar Favoritos Histórico Correio Imprimir

Endereço http://www.genopar.org/genome/annotation/index.php Ir Links

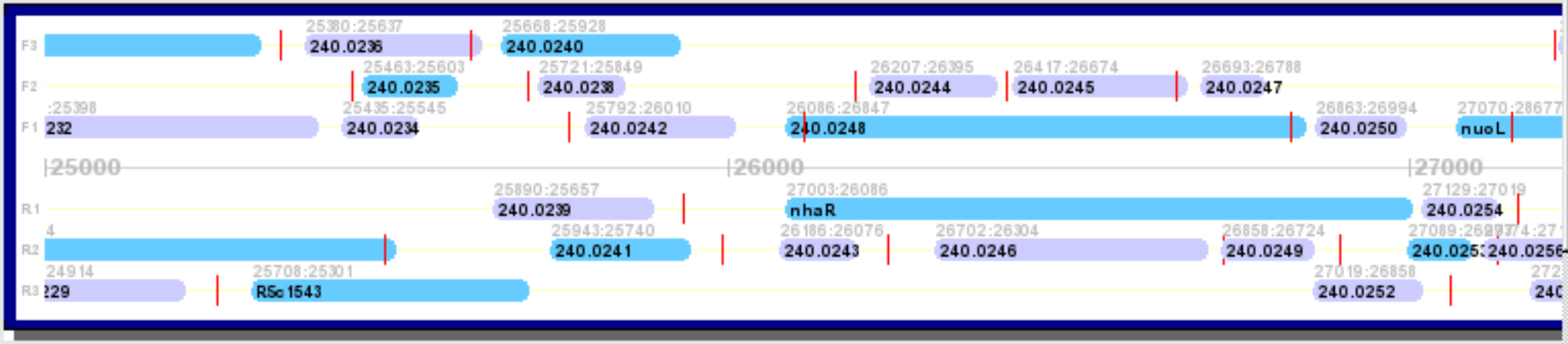
GENOPAR

Contig Annotation Page [MF lab]

Contig: 240 Show Contig

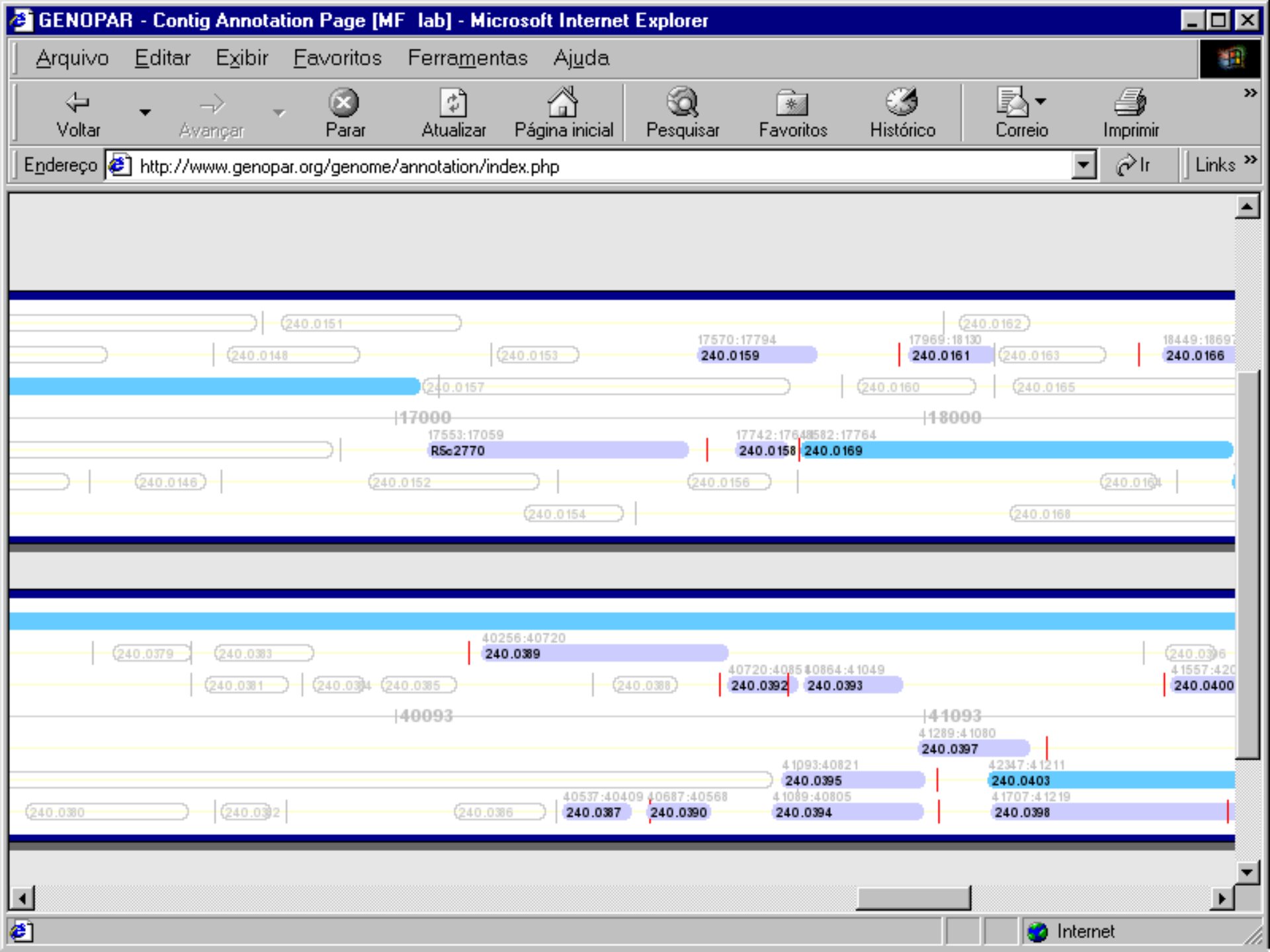
Start: 25000 Show Sequence Reverse

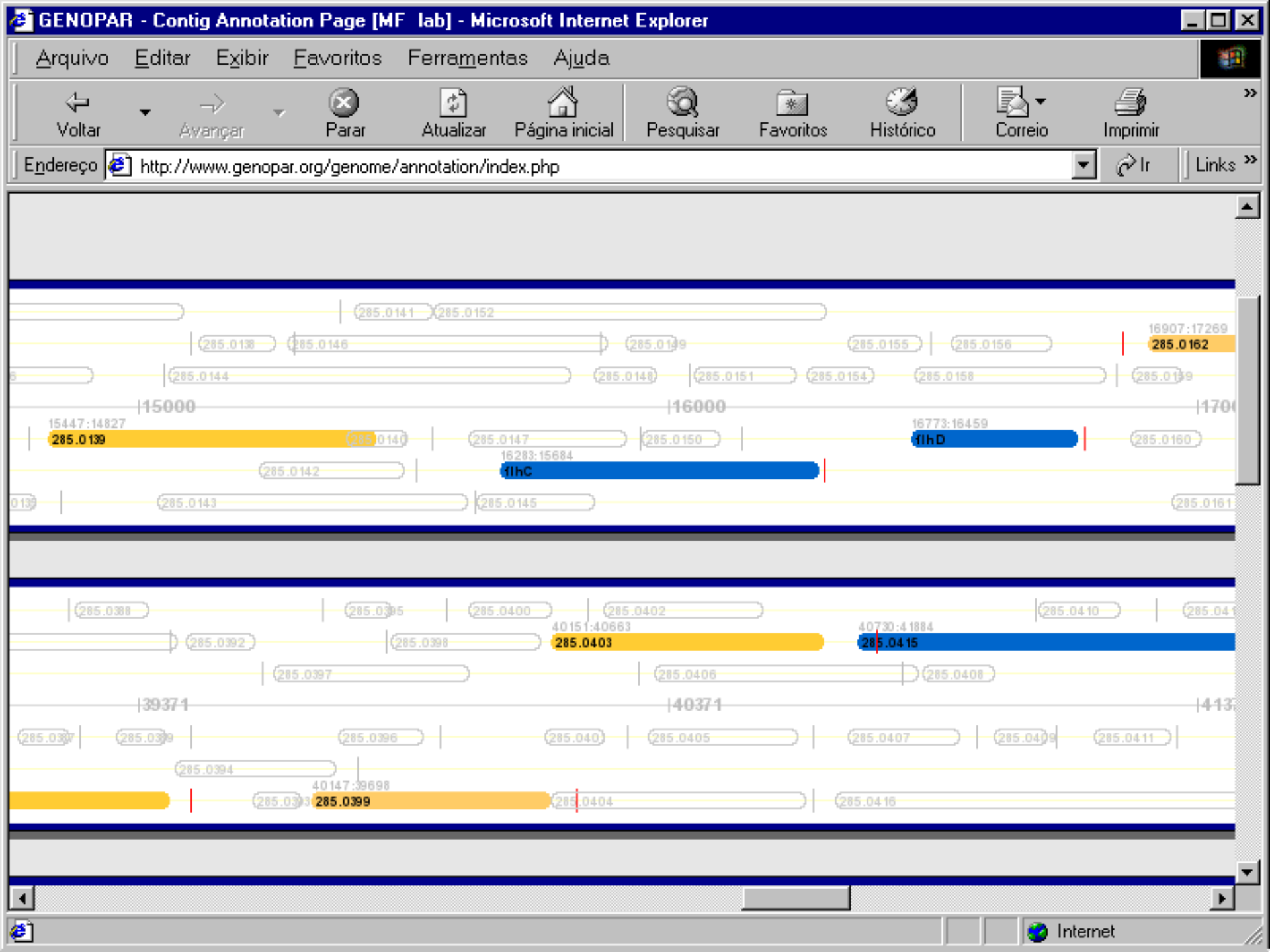
End: 40000 Show Contig Quality



http://www.genopar.org/genome/annotation/tmp/MAP240.25000.40000.3.map?456,101

Internet


GENOPAR - Contig Annotation Page [MF lab] - Microsoft Internet Explorer
Arquivo Editar Exibir Favoritos Ferramentas Ajuda
Voltar Avançar Parar Atualizar Página inicial Pesquisar Favoritos Histórico Correio Imprimir
Endereço http://www.genopar.org/genome/annotation/index.php
Ir Links
240.0151
240.0148
240.0153
17570:17794
240.0159
240.0162
17969:18130
240.0161
240.0163
18449:18697
240.0166
240.0157
240.0160
240.0165
17000
18000
17553:17059
RSc2770
17742:17764
240.0158
240.0169
240.0146
240.0152
240.0156
240.0154
240.0168
40256:40720
240.0389
240.0379
240.0383
240.0381
240.0385
240.0388
40720:40854
40864:41049
240.0392
240.0393
41557:42
240.0400
40093
41093
41289:41080
240.0397
41093:40821
240.0395
42347:41211
240.0403
240.0380
240.0386
40537:40409
40687:40568
240.0387
240.0390
41089:40805
240.0394
41707:41219
240.0398
Internet


GENOPAR - Contig Annotation Page [MF lab] - Microsoft Internet Explorer
Arquivo Editar Exibir Favoritos Ferramentas Ajuda
Voltar Avançar Parar Atualizar Página inicial Pesquisar Favoritos Histórico Correio Imprimir
Endereço http://www.genopar.org/genome/annotation/index.php
Ir Links
285.0141 285.0152
285.0138 285.0146 285.0149 285.0155 285.0156
16907:17269
285.0162
285.0144 285.0148 285.0151 285.0154 285.0158
15000 16000
15447:14827
285.0139
285.0140
285.0147 285.0150
16773:16459
flhD
285.0160
285.0142
16283:15684
flhC
285.0143 285.0145
285.0388 285.0400 285.0402 285.0410
40151:40663
285.0403
40730:41884
285.0415
285.0392 285.0398
285.0397 285.0406 285.0408
39371 40371
285.0396 285.040 285.0405 285.0407 285.0411
285.0394
40147:39698
285.0399
285.0404 285.0416
Internet

GenBank

- O NCBI, ou Centro Nacional para Informação Biotecnológica dos EUA, é considerado o banco de dados central sobre informações genômicas.
- Vários outros bancos de dados similares estão distribuídos por países da Europa e Japão, mas todos trocam dados em um intervalo de 24 horas com o NCBI.
- O GenBank é o principal banco de dados do NCBI e armazena todas sequencias disponíveis publicamente de DNA (de seqüências pequenas a genomas inteiros), RNA e proteínas.

David J. Lipman é uma figura muito respeitada no meio da Bioinformática. Lipman é o diretor do NCBI, um dos autores originais do BLAST, programa de alinhamento de sequência.

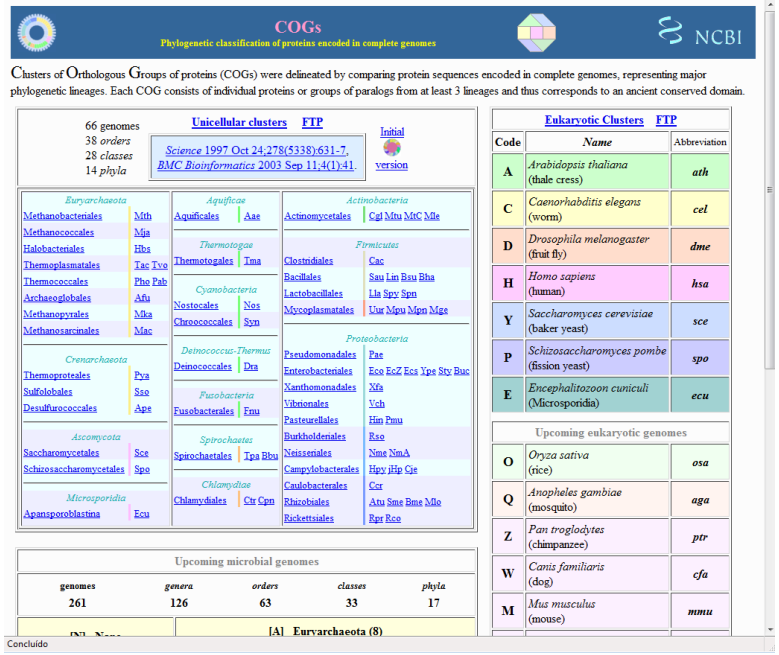
Fundado em 1988, o NCBI (Centro Nacional para Informação Biotecnológica) serve como um recurso nacional para informações sobre biologia molecular. Este é um órgão mantido pelo NIH (Institutos Nacionais de Saúde) dos Estados Unidos.

Bancos de dados secundários do NCBI

- **UniGene**
  - agrupa **todas as sequencias parciais do transcriptoma** de um organismo em aglomerados ou clusters, onde cada aglomerado representa a sequencia consenso de um gene.
- **Banco de dados RefSeq**
  - reúne somente as **sequencias de referência**, ou seja, a mais representativa sequencia de um transcrito, editada e inspecionada por um curador. É, frequentemente, o melhor banco de dados para se evitar a redundância natural num universo com tantas informações.
- **OMIM (Online Mendelian Inheritance in Man)**
  - que foi criado para **catalogar todos genes e alelos relacionados a doenças** e outras características humanas, bem como proporcionar um detalhamento técnico e bibliografia referente a cada característica.

COG

- Cluster of Orthologous Groups
  - 66 genomas bacterianos
- Best Hits cruzados entre 3 organismos
- Genes bacterianos agrupados por função biológica
- KOG, eucariotos



Página inicial NCBI

NCBI Resources How To

NCBI
National Center for Biotechnology Information

Search PubMed

PubMed
Protein
Nucleotide
GSS
EST
Structure
Genome
BioProject (Genome Project)
BioSample
BioSystems
Books
CancerChromosomes
Conserved Domains
dbGaP
dbVar
Epigenomics
Gene
GENSAT
GEO DataSets
GEO Profiles
HomoloGene
MeSH
NCBI Web Site
NLM Catalog
OMIA
OMIM
PMC
PopSet
Probe
Protein Clusters

Search Clear

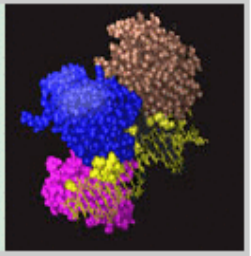
- NCBI Home
- Site Map (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

ology Information advances science and
medical and genomic information.

ization | Research | RSS Feeds

BI software
software
plish specific tasks at NCBI
enBank or other NCBI databases

res of pro-
Examine
active sites,
molecular interactions, biological activities of
bound chemicals, and associated biosystems.



II 1 2 3 4 5

Popular Resources

- BLAST
- Bookshelf
- Gene
- Genome
- Nucleotide
- OMIM
- Protein
- PubChem
- PubMed
- PubMed Central
- SNP

NCBI News

NCBI Discovery Workshop: A Practical Hands-On Course
02 Aug 2011
September 27-28, 2011 @ NLM: Space is still available in the 2-day

New NCBI News Issue
06 Jul 2011
Information on the redesigned PopSet resource, as well as new

More...

Display Settings: Summary, 20 per page, Sorted by Recently Added

Send to:

Results: 1 to 20 of 91

1. System responses to long-term drought and re-watering of two contrasting alfalfa varieties.
   Kang Y, Han Y, Torres-Jerez I, Wang M, Tang Y, Monteros M, Udvardi M.
   Plant J. 2011 Aug 12. doi: 10.1111/j.1365-313X.2011.04738.x. [Epub ahead of print]
   PMID: 21838776 [PubMed - as supplied by publisher]
   Related citations

2. Expression analysis of the gene family associated with raffinose accumulation in rice seedlings under cold stress.
   Saito M, Yoshida M.
   J Plant Physiol. 2011 Aug 6. [Epub ahead of print]
   PMID: 21824678 [PubMed - as supplied by publisher]
   Related citations

3. Expression of a **GALACTINOL** SYNTHASE gene is positively associated with desiccation tolerance of Brassica napus seeds during development.
   Li X, Zhuo J, Jing Y, Liu X, Wang X.
   J Plant Physiol. 2011 Oct 15;168(15):1761-1770. Epub 2011 Jun 15.
   PMID: 21680054 [PubMed - as supplied by publisher]
   Related citations

4. Functional characterization of a eukaryotic melibiose transporter.
   Lingner U, Münch S, Sode B, Deising HB, Sauer N.
   Plant Physiol. 2011 Jul;156(3):1565-76. Epub 2011 May 18.
   PMID: 21593216 [PubMed - in process] Free PMC Article
   Free full text Related citations

5. Comparative physiology and transcriptional networks underlying the heat shock response in Populus trichocarpa, Arabidopsis thaliana and Glycine max.
   Weston DJ, Karve AA, Gunter LE, Jawdy SS, Yang X, Allen SM, Wullschleger SD.

Filter your results:

All (91)
Free Full Text (57)
Review (0)

Manage Filters

Titles with your search terms

Galactinol and raffinose constitute a novel function to protect plants from oxidative [Plant Physiol. 2008]

Important roles of drought- and cold-inducible genes for **galactinol** synthase in stre [Plant J. 2002]

Galactinol is a signaling component of the induced systemic resistan [Mol Plant Microbe Interact. 2008]

See more...

41 free full-text articles in PubMed Central

Functional characterization of a eukaryotic melibiose transporter. [Plant Physiol. 2011]

Down-regulation of the myo-inositol oxygenase gene family has no effect on cell wall [Planta. 2011]

Interaction with diurnal and circadian regulation results in dynamic metabolic and [PLoS One. 2010]

See all (41)...

Find related data

Database: Select

NCBI

National Center for Biotechnology Information

Search All Databases

All Databases
PubMed
Protein
Nucleotide
GSS
EST
Structure
Genome
BioProject (Genome Project)
BioSample
BioSystems
Books
CancerChromosomes
Conserved Domains
dbGaP
dbVar
Epigenomics
Gene
GENSAT
GEO DataSets
GEO Profiles
HomoloGene
MeSH
NCBI Web Site
NLM Catalog
OMIA
OMIM
PMC
PopSet
Probe

Search Clear

- NCBI Home
- Site Map (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

ology Information advances science and
medical and genomic information.

ization | Research | RSS Feeds

BI software
software
mplish specific tasks at NCBI
GenBank or other NCBI databases

lp
ls, news
outlets, and other educational resources.

II 1 2 3 4 5

Popular Resources

- BLAST
- Bookshelf
- Gene
- Genome
- Nucleotide
- OMIM
- Protein
- PubChem
- PubMed
- PubMed Central
- SNP

NCBI News

NCBI Discovery Workshop: A Practical Hands-On Course
02 Aug 2011
September 27-28, 2011 @ NLM: Space is still available in the 2-day

New NCBI News Issue
06 Jul 2011
Information on the redesigned PopSet resource, as well as new

More...

Display Settings: Summary, 20 per page, Sorted by Default order Send to:

Found 236 nucleotide sequences. Nucleotide (220) EST (16)

Results: 1 to 20 of 220

<< First < Prev Page 1 of 11 Next > Last >>

1. Zea mays full-length cDNA clone ZM_BFb0208C05 mRNA, complete cds
   1,823 bp linear mRNA
   Accession: BT069688.1 GI: 224035018
   GenBank FASTA Graphics Related Sequences

2. Zea mays full-length cDNA clone ZM_BFc0055O11 mRNA, complete cds
   1,963 bp linear mRNA
   Accession: BT067455.1 GI: 224030552
   GenBank FASTA Graphics Related Sequences

3. Zea mays full-length cDNA clone ZM_BFc0157C17 mRNA, complete cds
   1,262 bp linear mRNA
   Accession: BT064294.1 GI: 223949814
   GenBank FASTA Graphics Related Sequences

4. Zea mays full-length cDNA clone ZM_BFc0043I06 mRNA, complete cds
   2,599 bp linear mRNA
   Accession: BT063253.1 GI: 223947732
   GenBank FASTA Graphics Related Sequences

5. Zea mays full-length cDNA clone ZM_BFb0204I06 mRNA, complete cds
   2,436 bp linear mRNA
   Accession: BT061992.1 GI: 223945210
   GenBank FASTA Graphics Related Sequences

6. Arabidopsis thaliana chromosome 5, complete sequence
   26,975,502 bp linear DNA

Filter your results:

All (220)
Bacteria (0)
INSDC (GenBank) (159)
mRNA (111)
RefSeq (61)

Manage Filters

Top Organisms [Tree]

Arabidopsis thaliana *(50)*
Sorghum bicolor *(17)*
Laccaria bicolor S238N-H82 *(12)*
Zea mays *(10)*
Cucumis melo *(6)*
All other taxa *(64)*
More...

Find related data

Database: Select

Find items

Search details

```
galactinol[All Fields] AND
plant[All Fields]
```

Display Settings: GenBank Send:

Zea mays full-length cDNA clone ZM_BFc0157C17 mRNA, complete cds

GenBank: BT064294.1

FASTA Graphics

Go to:

```
LOCUS       BT064294               1262 bp    mRNA    linear   PLN 21-FEB-2009
DEFINITION  Zea mays full-length cDNA clone ZM_BFc0157C17 mRNA, complete cds.
ACCESSION   BT064294
VERSION     BT064294.1  GI:223949814
KEYWORDS    FLI_CDNA.
SOURCE      Zea mays
  ORGANISM  Zea mays
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; PACMAD
            clade; Panicoideae; Andropogoneae; Zea.
REFERENCE   1  (bases 1 to 1262)
  AUTHORS   Yu,Y., Currie,J., Lomeli,R., Angelova,A., Collura,K., Wissotski,M.,
            Campos,D., Kudrna,D., Golser,W., Ashely,E., Haller,K., Descour,A.,
            Fernandes,J., Soderlund,C. and Walbot,V.
  TITLE     Maize Full-length cDNA Project
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1262)
  AUTHORS   Yu,Y., Currie,J., Lomeli,R., Angelova,A., Collura,K., Wissotski,M.,
            Campos,D., Kudrna,D., Golser,W., Ashely,E., Haller,K., Descour,A.,
            Fernandes,J., Soderlund,C. and Walbot,V.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-FEB-2009) Plant Sciences, Arizona Genomics Institute,
            1657 E. Helen Street, Thomas Keating Building, Tucson, AZ 85721,
            USA
FEATURES             Location/Qualifiers
     source          1..1262
                     /organism="Zea mays"
                     /mol_type="mRNA"
```

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Find in this Sequence

LinkOut to external resources

Gramene

[Gramene]

Related information

Related Sequences

Protein

Taxonomy

UniGene

Recent activity

Turn Off Clear

Zea mays full-length cDNA clone ZM_BFc0157C17 mRNA, complete cd: Nucleotide

galactinol plant (220)

Nucleotide

Display Settings: FASTA Send:

Zea mays full-length cDNA clone ZM_BFc0157C17 mRNA, complete cds

GenBank: BT064294.1

GenBank Graphics

>gi|223949814|gb|BT064294.1| Zea mays full-length cDNA clone ZM_BFc0157C17 mRNA, complete cds

GGCCGCGTCGCCATCACCCGCGCCTACCACCGCGCGCTCGAGGCCTCCGTGGCGCGCAGCTTCCCGGACA
ACGGCTGCATCTCCTGCATGTGCCACAACTCCGACATGCTCTACAGCGCCAGGCAGACTGCCGTCGTGCG
CGCCTCCGACGACTTCTACCCGCGCGACCCGGCATCGCACACCGTCCACGTCGCCTCCGTCGCGTACAAC
ACCGTCTTCCTCGGCGAGTTCATGCAGCCCGATTGGGACATGTTCCATAGCTTGCATCCGGCGGCGGAGT
ACCACGGCGCGGCGAGGGCCATCGGTGGCTGCCCGATATACGTCAGCGACAAGCCGGGGAACCACAACTT
CGAGCTGCTCAGGAAGCTCGTGCTCCCCGACGGCACCGTGCTACGCGCGCAGCTTCCCGGCCGGCCCACA
CGGGACTGCCTCTTCTCCGACCCGGCGCGCGACGGCGCGAGTTTGCTCAAGATTTGGAACCTGAACAAGT
GCGGTGGCGTGGTGGGTGTGTTCAACTGCCAGGGAGCCGGGTGGTGCCGCGTGACCAAGCGGACGCGCGT
GCACGACGCGTCGCCGGGCACGCTGACCGGCACCGTGCGTGCCGACGACGTCGACGCCATAGCGCGCGTC
GCTGGTGACGGCGGCGGGTGGGACGGCGAGACCGTGGTGTATGCGCACCGCACGCGGGAGCTAGTGCGAC
TGCCCCGGGGCGTCGCGCTGCCTGTGACGCTAGGCCCGCTCCAGTATGAGGTGTTCCATGTGTGCCCGCT
CCGCGCCGTCGTGCCGGGGTTCTCGTTCGCGCCCGTCGGGCTGCTCGATATGTTCAACGCTGGGGGCGCC
GTTGAGGAGTGCGACGTGATCAGCAATGTCGGCGGCAAGGCCATGGCTCTCAGGGTTCGCGGGTGCGGTC
GGTTCGGCGCTTACTGCTCGCGGGAGCCGGCGAGGTGTCTATTGGACTCGGCGGAAGTGGAGTTCAGCTA
CGATGCCGACACCGGCCTCGTGTCCGTCGACCTGCCCGTGCCGGAGCAGGAGCTATATCGGTGGACGCTG
GAGATTATGGTCTAGGCCGTGACGCTGTGCGCTGACGCCATTCTGAGCCGGACGACATTGTCATTTGCTA
TGTGCACGAGAGATAGATGTATTGTGGTGTTGTTGTATAATTGGTATGGAGAGAGTATGCTTTTGTTATG
TCTGTATCTGCATACCCTGCGTGCTTTGAATGTTATATAATACTTGATAGATAATAAAGGATGGTGTTCC
TG

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Find in this Sequence

LinkOut to external resources

Gramene

[Gramene]

Related information

Related Sequences

Protein

Taxonomy

UniGene

Recent activity

Turn Off Clear

Zea mays full-length cDNA clone ZM_BFc0157C17 mRNA, complete cd: Nucleotide

galactinol plant (220)

Nucleotide

ORF Finder (Open Reading Frame Finder)

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database.
This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

Enter GI or ACCESSION OrfFind Clear

or sequence in FASTA format

FROM: **TO:**

Genetic codes 1 Standard

Comments and suggestions to: info@ncbi.nlm.nih.gov
Credits to: Tatiana Tatusov and Roman Tatusov

ORF Finder (Open Reading Frame Finder)

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database.
This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

Enter GI or ACCESSION OrfFind Clear

or sequence in FASTA format

```
AGTATGTCCAGCATTG
GTGCAAGTCAATCCTATGCTCTACCGTATACTTTGGCCACATATGTAGTTCAAT
GCTGGTTTGTATCACA
ATCAAATAAATATCTAGCACTGAAAAGGAGCGTGTAATTTCTCTGTAATTCACC
AATAAATCGCAATCAA
ATAATCTCTTTCTTAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAA
AAAAAAAAAAAA
```

FROM: **TO:**

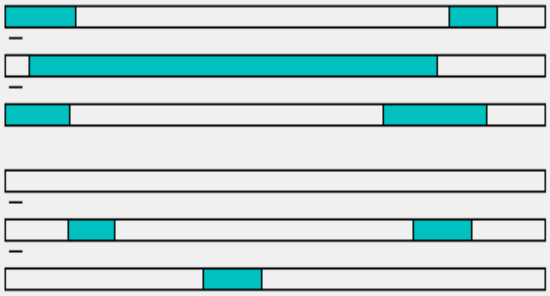
Genetic codes 1 Standard

Comments and suggestions to: info@ncbi.nlm.nih.gov
Credits to: Tatiana Tatusov and Roman Tatusov

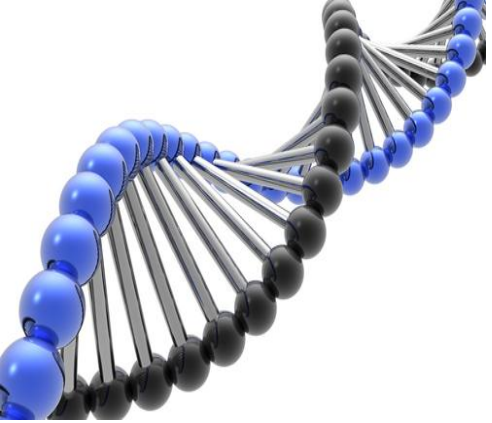
ORF Finder (Open Reading Frame Finder)

Anonymous

View 1 GenBank Redraw 100 SixFrames



Frame	from	to	Length
+2	62	1078	1017
+3	942	1199	258
+1	1	177	177
+3	3	164	162
-2	1016	1162	147
-3	496	639	144
+1	1108	1227	120
-2	161	277	117



DNA codifica 6 proteínas potenciais

5′ CAT CAA

5′ ATC AAC

5′ TCA ACT

5′ CATCAACTACAACTCCAAAGACACCCTTACACATCAACAAACCTACCCAC 3′

3′ GTAGTTGATGTTGAGGTTTCTGTGGGAATGTGTAGTTGTTTGGATGGGTG 5′

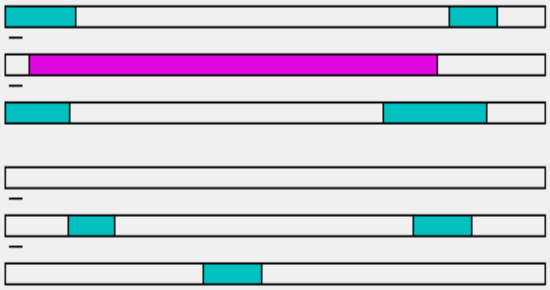
5′ GTG GGT

5′ TGG GTA

5′ GGG TAG

Program blastp Database nr BLAST with parameters Cognitor

View 1 GenBank Redraw 100 SixFrames



Length: 338 aa

Accept Alternative Initiation Codons

Frame	from	to	Length
+2	62	1078	1017
+3	942	1199	258
+1	1	177	177
+3	3	164	162
-2	1016	1162	147
-3	496	639	144
+1	1108	1227	120
-2	161	277	117

```
 62 atggcccctcaagaagtacccgcagaagccttcaccgccgcaggg
    M  A  P  Q  E  V  P  A  E  A  F  T  A  A  G
107 aaagtctcaacactcagcaacacaggctactcgaagagagcgtac
    K  V  S  T  L  S  N  T  G  Y  S  K  R  A  Y
152 gttacattttagctggaaatggtgactacgtgaaaggagtggtt
    V  T  F  L  A  G  N  G  D  Y  V  K  G  V  V
197 ggattggctaaaggtctgaggaaggttaacagtgcctacccccttt
    G  L  A  K  G  L  R  K  V  N  S  A  Y  P  L
242 gtggttgccatcctcccggatgttcccgaggaacatcgggaaata
    V  V  A  I  L  P  D  V  P  E  E  H  R  E  I
287 ttaaggtcacagggttgcattgttcgtgaaattgagcccatttac
    L  R  S  Q  G  C  I  V  R  E  I  E  P  I  Y
332 ccacctgaaaaccagattcagtttgctatggcttactatgtcatc
    P  P  E  N  Q  I  Q  F  A  M  A  Y  Y  V  I
377 aattattccaagctccgcatatggaactttgaggagtacagcaag
    N  Y  S  K  L  R  I  W  N  F  E  E  Y  S  K
422 atgatctacctagacgcagacatccaagtttatgacaacatcgac
    M  I  Y  L  D  A  D  I  Q  V  Y  D  N  I  D
467 catctgtttgacgcagcagatggatactttatgcagtgatggat
    H  L  F  D  A  A  D  G  Y  F  Y  A  V  M  D
512 tgcttctgcgagaagacttggagtaattccccacagtactccatt
    C  F  C  E  K  T  W  S  N  S  P  Q  Y  S  I
557 ggatattgccaacagtgcccggacaaagttacctggccagctgat
```

>lcl|Sequence 1 ORF:62..1078 Frame +2
MAPQEVPAEAFTAAGKVSTLSNTGYSKRAYVTFLAGNGDYVKGVVGLAKGLRKVNSAYPLVVAILPDVPE
EHREILRSQGCIVREIEPIYPPENQIQFAMAYYVINYSKLRIWNFEEYSKMIYLDADIQVYDNIDHLFDA
ADGYFYAVMDCFCEKTWSNSPQYSIGYCQQCPDKVTWPADMGSPPPLYFNAGMFVFEPSRLTYENLLETL
QITPPTLFAEQDFLNMFFQTTYKPISLAYNLVLAMLWRHPENVELDEVKVVHYCAAGSKPWRYTGKEANM
DREDIKMLVQKWWDVYNDASLDFKAEDPVPEEETFSRPSVMAFMPEPAISYVPAPSAA*

5. Tuli-Raffinose synthase 3
783 aa protein
Accession: Q9FND9.1 GI: 75171832
GenPept FASTA Graphics Related Sequences Identical Proteins

galactinol synthase 1 [Arabidopsis thaliana]
6. 344 aa protein
Accession: AEC10811.1 GI: 330255717
GenPept FASTA Graphics Related Sequences Identical Proteins

putative **galactinol**--sucrose galactosyltransferase 5 [Arabidopsis thaliana]
7. 783 aa protein
Accession: AED94542.1 GI: 332007159
GenPept FASTA Graphics Related Sequences Identical Proteins

galactinol synthase 4 [Arabidopsis thaliana]
8. 334 aa protein
Accession: AEE33692.1 GI: 332195571
GenPept FASTA Graphics Related Sequences Identical Proteins

galactinol synthase 7 [Arabidopsis thaliana]
9. 332 aa protein
Accession: AEE33688.1 GI: 332195567
GenPept FASTA Graphics Related Sequences Identical Proteins

galactinol synthase 2 [Arabidopsis thaliana]
10. 335 aa protein
Accession: AEE33413.1 GI: 332195292
GenPept FASTA Graphics Related Sequences Identical Proteins

galactinol synthase 3 [Arabidopsis thaliana]
11. 334 aa protein
Accession: AEE28432.1 GI: 332190311
GenPept FASTA Graphics Related Sequences Identical Proteins

galactinol synthase 1 [Arabidopsis thaliana]
12. 344 aa protein
Accession: NP_182240.1 GI: 15226522

plant[All Fields]

Search

See more...

Recent activity

Turn Off Clear

galactinol plant (211) Protein

galactinol palnt (0) Protein

galactinol (916) Protein

Zea mays full-length cDNA clone ZM_BFc0157C17 mRNA, complete cd: Nucleotide

galactinol plant (220) Nucleotide

See more...

Display Settings: GenPept Send to:

galactinol synthase 2 [Arabidopsis thaliana]

GenBank: AEE33413.1

FASTA Graphics

Go to:

```
LOCUS       AEE33413               335 aa            linear   PLN 09-JUN-2011
DEFINITION  galactinol synthase 2 [Arabidopsis thaliana].
ACCESSION   AEE33413
VERSION     AEE33413.1  GI:332195292
DBLINK      Project: 10719
DBSOURCE    accession CP002684.1
KEYWORDS    .
SOURCE      Arabidopsis thaliana (thale cress)
  ORGANISM  Arabidopsis thaliana
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
            rosids; malvids; Brassicales; Brassicaceae; Camelineae;
            Arabidopsis.
REFERENCE   1  (residues 1 to 335)
  AUTHORS   Theologis,A., Ecker,J.R., Palm,C.J., Federspiel,N.A., Kaul,S.,
            White,O., Alonso,J., Altafi,H., Araujo,R., Bowman,C.L.,
            Brooks,S.Y., Buehler,E., Chan,A., Chao,Q., Chen,H., Cheuk,R.F.,
            Chin,C.W., Chung,M.K., Conn,L., Conway,A.B., Conway,A.R.,
            Creasy,T.H., Dewar,K., Dunn,P., Etgu,P., Feldblyum,T.V., Feng,J.,
            Fong,B., Fujii,C.Y., Gill,J.E., Goldsmith,A.D., Haas,B.,
            Hansen,N.F., Hughes,B., Huizar,L., Hunter,J.L., Jenkins,J.,
            Johnson-Hopson,C., Khan,S., Khaykin,E., Kim,C.J., Koo,H.L.,
            Kremenetskaia,I., Kurtz,D.B., Kwan,A., Lam,B., Langin-Hooper,S.,
            Lee,A., Lee,J.M., Lenz,C.A., Li,J.H., Li,Y., Lin,X., Liu,S.X.,
            Liu,Z.A., Luros,J.S., Maiti,R., Marziali,A., Militscher,J.,
            Miranda,M., Nguyen,M., Nierman,W.C., Osborne,B.I., Pai,G.
```

Change region shown

Customize view

Analyze this sequence

Run BLAST

Identify Conserved Domains

Find in this Sequence

Articles about the GolS2 gene

Metabolic pathways involved in cold acclimation identified by integrated analysi [Plant Physiol. 2009]

Galactinol and raffinose constitute a novel function to protect plants from oxidative [Plant Physiol. 2008]

See all...

Identical proteins for AEE33413.1

Sequence 10 from patent US 7807406 [ADS59960]

Sequence 8 from patent US 7476778 [ACN00706]

Sequence 8 from patent US 7294756 [ABY02571]

See all...

Reference sequence information

RefSeq protein

Display Settings: FASTA Send to:

galactinol synthase 2 [Arabidopsis thaliana]

GenBank: AEE33413.1

GenPept Graphics

```
>gi|332195292|gb|AEE33413.1| galactinol synthase 2 [Arabidopsis thaliana]
MAPEINTKLTVPVHSATGGEKRAYVTFLAGTGDYVKGVVGLAKGLRKAKSKYPLVVAVLPDVPEDHRKQL
VDQGCVVKEIEPVYPPENQTEFAMAYYVINYSKLRIWEFVEYNKMIYLDGDIQVFDNIDHLFDLPNGQFY
AVMDCFCEKTWSHSPQYKIGYCQQCPDKVTWPEAKLGPKPPLYFNAGMFVYEPNLSTYHNLLETVKIVPP
TLFAEQDFLNMYFKDIYKPIPPVYNLVLAMLWRHPENIELDQVKVVHYCAAGAKPWRFTGEEENMDREDI
KMLVKKWWDIYNDESLDYKNVVIGDSHKKQQTLQQFIEALSEAGALQYVKAPSAA
```

Change region shown

Analyze this sequence

Run BLAST

Identify Conserved Domains

Find in this Sequence

Articles about the GolS2 gene

Metabolic pathways involved in cold acclimation identified by integrated analysis [Plant Physiol. 2009]

Galactinol and raffinose constitute a novel function to protect plants from oxidative [Plant Physiol. 2008]

See all...

Identical proteins for AEE33413.1

Sequence 10 from patent US 7807406 [ADS59960]

Sequence 8 from patent US 7476778 [ACN00706]

Sequence 8 from patent US 7294756 [ABY02571]

See all...

Reference sequence information

RefSeq protein

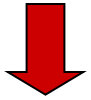
See the reference protein sequence for galactinol synthase 2 (NP_176053.1).

Para comparar é necessário alinhar...

Duas sequências são combinadas aleatoriamente

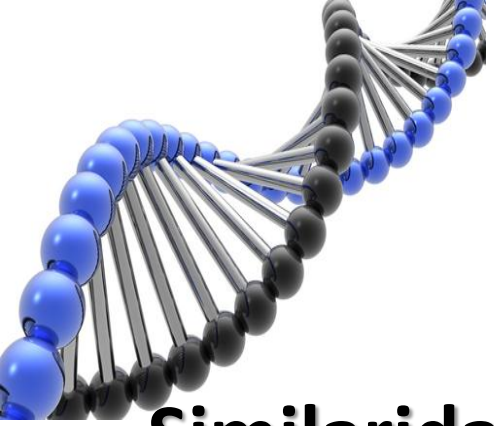
Qualidade da combinação é avaliada e pontuada

Uma sequência é movida em relação a outra e a combinação é pontuada novamente



Obtenção da melhor pontuação de alinhamento

Método automatizado → Alinhamento entre as milhares de alternativas



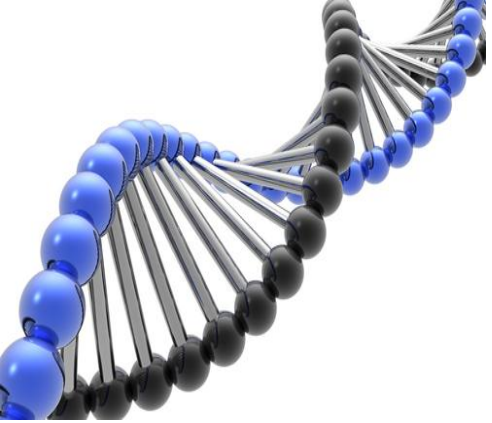
Conceitos importantes

Similaridade : é uma medida da qualidade do alinhamento entre as sequencias, baseada em algum critério de comparação.

Homólogos: genes que descendem de um ancestral comum (por ex., todas as globinas);

Parálogos: homólogos que divergiram depois de duplicação génica (por ex., α e β-globina);

Ortólogos: homólogos que divergiram devido à especiação (por ex., β-globinas de humano e chimpanzé);



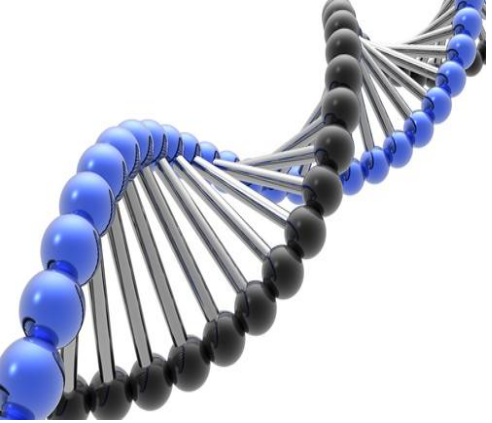
SIMILARIDADE
(valores estatísticos)

↓

HOMOLOGIA
(CONCLUSÃO)

Como definir função?

- Alinhamento de sequências
- Motivos (padrões consensuais)
- Blocos, perfis, etc....
- Hidden Markov Models - HMM



Alinhamento

```
0  1 2  3  4 5  6 7 8 9 10
G  A A  -  G G  A T T A G
G  A T  C  G G  A -  - A G
```

- **Identidade - MATCH**
- **Semelhança / divergência - MISMATCH**
- **Lacunas - GAPS**
- **Inserção/Deleção - INDELS**

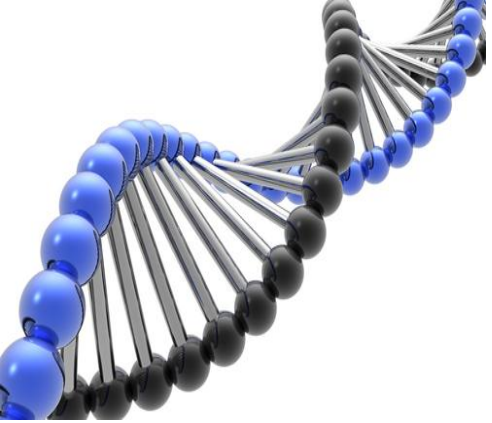
Global *vs.* Local

Global

- **Alinhamento de toda a sequência utilizado o maior número de caracteres possíveis**
- **Sequências similares e de tamanho aproximado**

Local

- **Segmentos com o maior número de identidades**
- **Regiões alinhadas e não alinhadas (≠ mismatch)**
- **Sequências similares em algumas regiões, que diferem em tamanho ou que compartilham domínios conservados**



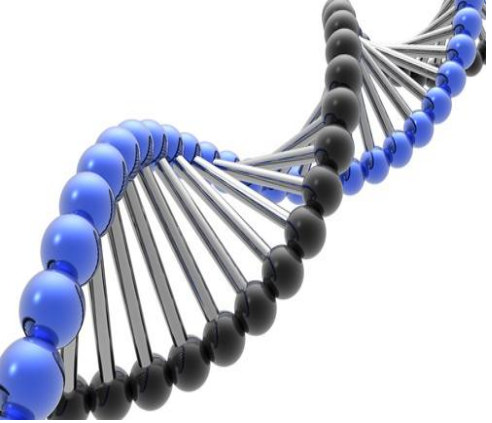
Aplicações

Global

- **Deduzir histórias evolutivas entre membros da mesma família**
- **Estabelecer a existência de um ancestral comum (homologia)**

Local

- **Inferir funções biológicas**
- **Identificar regiões conservadas e de alta similaridade (sítio ativo, domínios) entre outras pouco conservadas**
- **Reconstruir sequências de DNA a partir de seus fragmentos**
- **Comparar sequências de mRNA (sem íntrons) à sequência genômica**



Scores e Estatística

E-value

- **Significado estatístico do alinhamento**
- **Quanto menor o escore, mais significativo é o alinhamento**
- **E-value = 0.05. Significa que existem 5 chances em 100 (1 em 20) da similaridade entre as sequências ocorrer aleatoriamente**
- **Influenciado pelo tamanho do banco de dados e o sistema de escore utilizado**

BLAST Assembled Genomes

Choose a species genome to search, or **list all genomic BLAST databases**.

- **Human**
- **Mouse**
- **Rat**
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

Modalidades do BLAST

- **tBLASTx** foi utilizado em descoberta gênica inúmeras vezes, como por exemplo na identificação da subunidade catalítica da telomerase humana assim que tal enzima foi identificada no protozoário *Euplotes* (Meyerson et al. 1997).
- **BLASTn,** buscam homologia entre sequencias de nucleotídeos
- **BLASTp**, buscam homologia entre sequencias de proteínas
- **BLASTx**, Buscam homologia entre sequencias de nucleotídeos e proteínas
- **PSI-BLAST**, que em uma **primeira busca encontra as proteínas mais homólogas à pesquisada - Query**; procede identificando as regiões conservadas dentre os melhores resultados da pesquisa e**, em buscas subsequentes, mascara as regiões não conservadas da *Query* e pesquisa levando em conta apenas as regiões conservadas.**

Modalidades do BLAST

- **Input (Entrada do Programa)**
  - Query sequence (sequencia de busca)
  - Subject (Banco de dados de sequencias biológicas)
- **Output (Saída do Programa)**
  - Uma lista ordenada de “hits” contendo sequencias do banco de dados que possuem similaridade local com a
  - sequencia de busca (da qual a função desconhecida da sequencia de busca pode ser inferida).
  - Significância estatística de cada “hit”

NCBI/ **BLAST Home**

BLAST finds regions of similarity between biological sequences. more...

New Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool. Go

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

News

SOAP BLAST

A SOAP based BLAST service is available.
Mon, 18 Jul 2011 08:00:00 EST

More BLAST news...

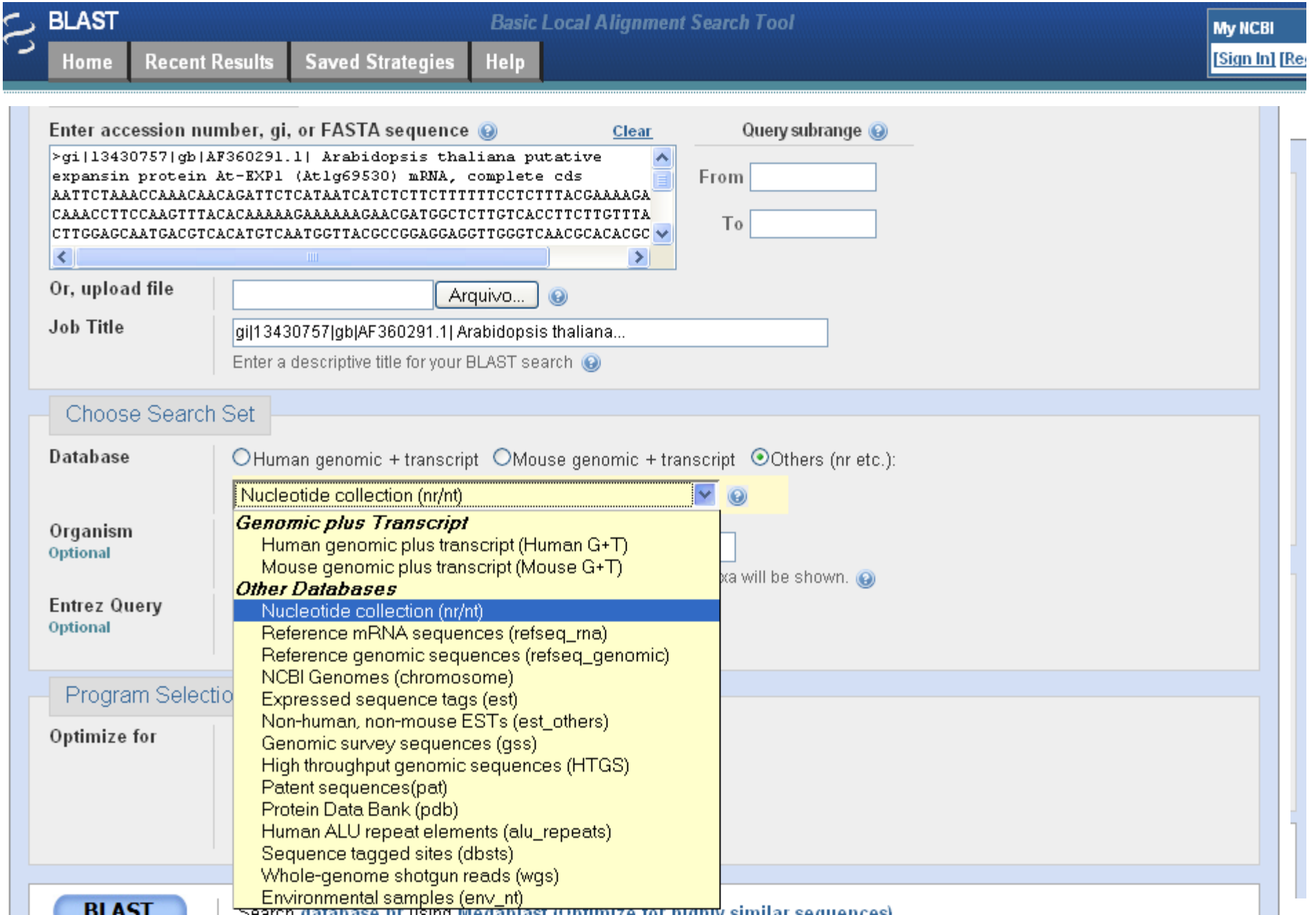
Tip of the Day

Use Genomic BLAST to see the genomic context

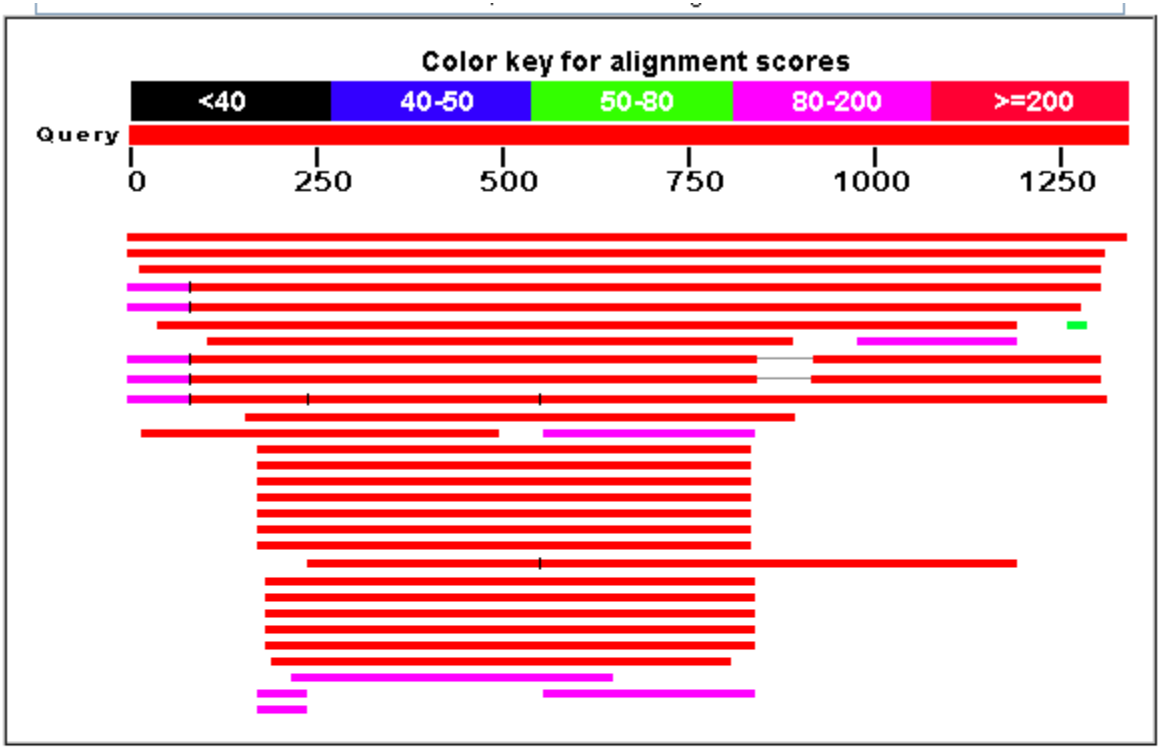
If you are interested in the evolution of a particular gene or gene family it is often intetesting to examine the intro-exon structure even across species.

More tips...

Blast N


BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In]
Enter accession number, gi, or FASTA sequence
Clear
>gi|13430757|gb|AF360291.1| Arabidopsis thaliana putative
expansin protein At-EXP1 (At1g69530) mRNA, complete cds
AATTCTAAACCAAACAACAGATTCTCATAATCATCTCTTCTTTTTTCCTCTTTACGAAAAGA
CAAACCTTCCAAGTTTACACAAAAAGAAAAAAGAACGATGGCTCTTGTCACCTTCTTGTTTA
CTTGGAGCAATGACGTCACATGTCAATGGTTACGCCGGAGGAGGTTGGGTCAACGCACACGC
Query subrange
From
To
Or, upload file
Arquivo...
Job Title
gi|13430757|gb|AF360291.1| Arabidopsis thaliana...
Enter a descriptive title for your BLAST search
Choose Search Set
Database
Human genomic + transcript
Mouse genomic + transcript
Others (nr etc.):
Nucleotide collection (nr/nt)
Genomic plus Transcript
Human genomic plus transcript (Human G+T)
Mouse genomic plus transcript (Mouse G+T)
Other Databases
Nucleotide collection (nr/nt)
Reference mRNA sequences (refseq_rna)
Reference genomic sequences (refseq_genomic)
NCBI Genomes (chromosome)
Expressed sequence tags (est)
Non-human, non-mouse ESTs (est_others)
Genomic survey sequences (gss)
High throughput genomic sequences (HTGS)
Patent sequences(pat)
Protein Data Bank (pdb)
Human ALU repeat elements (alu_repeats)
Sequence tagged sites (dbsts)
Whole-genome shotgun reads (wgs)
Environmental samples (env_nt)
Organism
Optional
Entrez Query
Optional
Program Selectio
Optimize for
BLAST

Alinhamento Blast N



Distance tree of results NEW

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
AF360291.1	Arabidopsis thaliana putative expansin protein At-EXP1 (At1g	2473	2473	100%	0.0	100%	U E G
NM_179537.3	Arabidopsis thaliana ATEXPA1 (ARABIDOPSIS THALIANA EXP	2418	2418	97%	0.0	100%	U G
BX815073.1	Arabidopsis thaliana Full-length cDNA Complete sequence frc	2338	2338	96%	0.0	99%	U G

>gb|AF229437.1| Arabidopsis thaliana expansin 10 (EXPA10) mRNA, complete cds
Length=1189

GENE ID: 839218 ATEXPA10 | ATEXPA10 (ARABIDOPSIS THALIANA EXPANSIN A10)
[Arabidopsis thaliana] (10 or fewer PubMed links)

Score = 544 bits (294), Expect = 5e-151
Identities = 548/670 (81%), Gaps = 20/670 (2%)
Strand=Plus/Plus

```
Query  176  CGGAGGAGGTTGGGTCAACGCACACGCCACATTCTACGGTGGTGGTGATGCTTCCGGCAC  235
            ||| || |||||| ||||||| |||||||| || ||||||||||||||||||||||||||
Sbjct  109  CGGTGGCGGTTGGATCAACGCTCACGCCACTTTTTACGGTGGTGGTGATGCTTCCGGCAC  168

Query  236  AATGGGAGGTGCTTGTGGATACGGAAACCTATATAGCCAAGGCTATGGAACCAACACGGC  295
            |||||| |||||||||||||| || || |||||||||||||||| || || | ||||||
Sbjct  169  AATGGGTGGTGCTTGTGGATATGGTAATCTATATAGCCAAGGCTACGGGACGAGCACGGC  228

Query  296  GGCGCTAAGCACGGCTCTATTCAATAATGGTCTAAGTTGTGGTGCTTGCTTCGAGATAAG  355
            ||| |||||||| ||||| ||||| ||||| || || |||||| ||||||| ||||||||
Sbjct  229  GGCTCTAAGCACAGCTCTCTTCAACAATGGACTTAGCTGTGGTTCTTGCTTTGAGATAAG  288

Query  356  ATGTCAAAACGATGGAAAATGGTGTCTT-CCTGGCTCAATTGTCGTCACAGCCACAAACT  414
            |||| |||||||||| ||||||||| || |||||||||| || || || || |||||||
Sbjct  289  ATGTGAAAACGATGGTAAATGGTGT-TTACCTGGCTCAATCGTTGTAACCGCTACAAACT  347

Query  415  TTTGCCCTCCTAACAACGCCTTACCGAACAACGCA-GGAGGTTGGTGTAACCCTCCTCA-  472
            | ||||| || || ||||| ||| |||||||   | || |||||||||| |||||||
Sbjct  348  TCTGCCCGCCAAATAACGCGTTAGCGAACAATA-ATGGCGGTTGGTGTAATCCTCCTCTT  406

Query  473  GCAGCATTTTGATCTCT-CTCAGCCCGTATTTCAACGCATCGCTCAATACAGAGCCGGCA  531
            | | || ||||| || | ||||||| || |||||||||| ||||| |||||||| || |
Sbjct  407  GAA-CACTTTGACCT-TGCTCAGCCTGTTTTTCAACGCATTGCTCAGTACAGAGCTGGAA  464

Query  532  TTGTCCCCGTCGCTTACCGAAGAGTGCCGTGC-GTGAGAAGAGGAGGAATAAGGTTTACG  590
            | ||||| ||  | ||| |||| || || ||| | ||||||||||||||||| || |||
Sbjct  465  TCGTCCCTGTTTCCTACAGAAGGGTTCCTTGCAG-GAGAAGAGGAGGAATAAGATTCACG  523

Query  591  ATAAACGGACACTCTTACTTCAACCTAGTTCTGATCACTAACGTCGGAGGAGCCGGAGAT  650
            |||||||| ||||| |||||||||| || |||||||| |||||||| || ||||||||
Sbjct  524  ATAAACGGCCACTCATACTTCAACCTTGTGCTGATCACAAACGTCGGTGGTGCCGGAGAC  583

Query  651  GTTCACTCAGCGATGGTTAAAGGTTCAAGAACTG-GATGGCAAGCGATGTCAAGAAACTG  709
            |||||||| |||  | | || |||||||||| | | |||||||| |||||||| |||||
Sbjct  584  GTTCACTCGGCGGCGATCAAGGGTTCAAGAACAGTG-TGGCAAGCTATGTCAAGGAACTG  642

Query  710  GGGACAGAACTGGCAGAGTAACTCTTACCTTAACGGACAATCTCTCTCATTCAAAGTTAC  769
            ||| || || ||||| || |||||||||| ||||| ||| | || || || || || ||
Sbjct  643  GGGGCAAAATTGGCAAAGCAACTCTTACCTCAACGGTCAAGCACTTTCCTTTAAGGTCAC  702
```

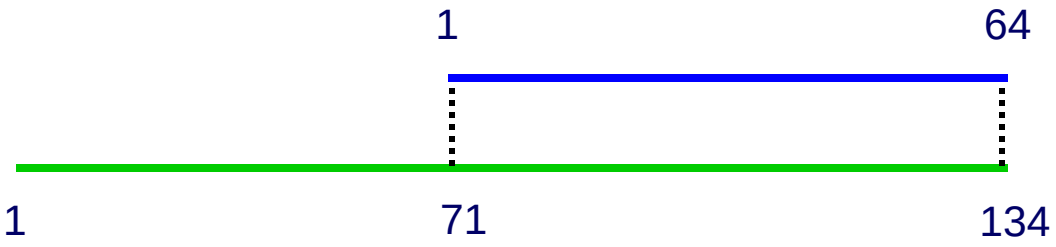
Alignments

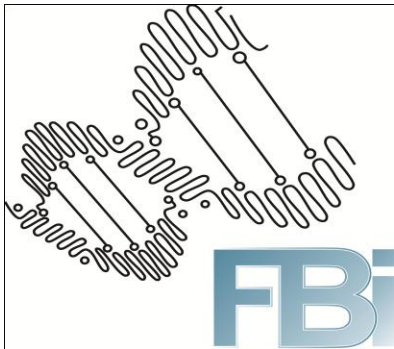
>gi|3913700|sp|Q52083|FUR_PSEPU FERRIC UPTAKE REGULATION PROTEIN (FERRIC UPTAKE REGULATOR)
Length = 134

Score = 79.7 bits (195), Expect = 9e-16
Identities = 42/64 (65%), Positives = 52/64 (80%)

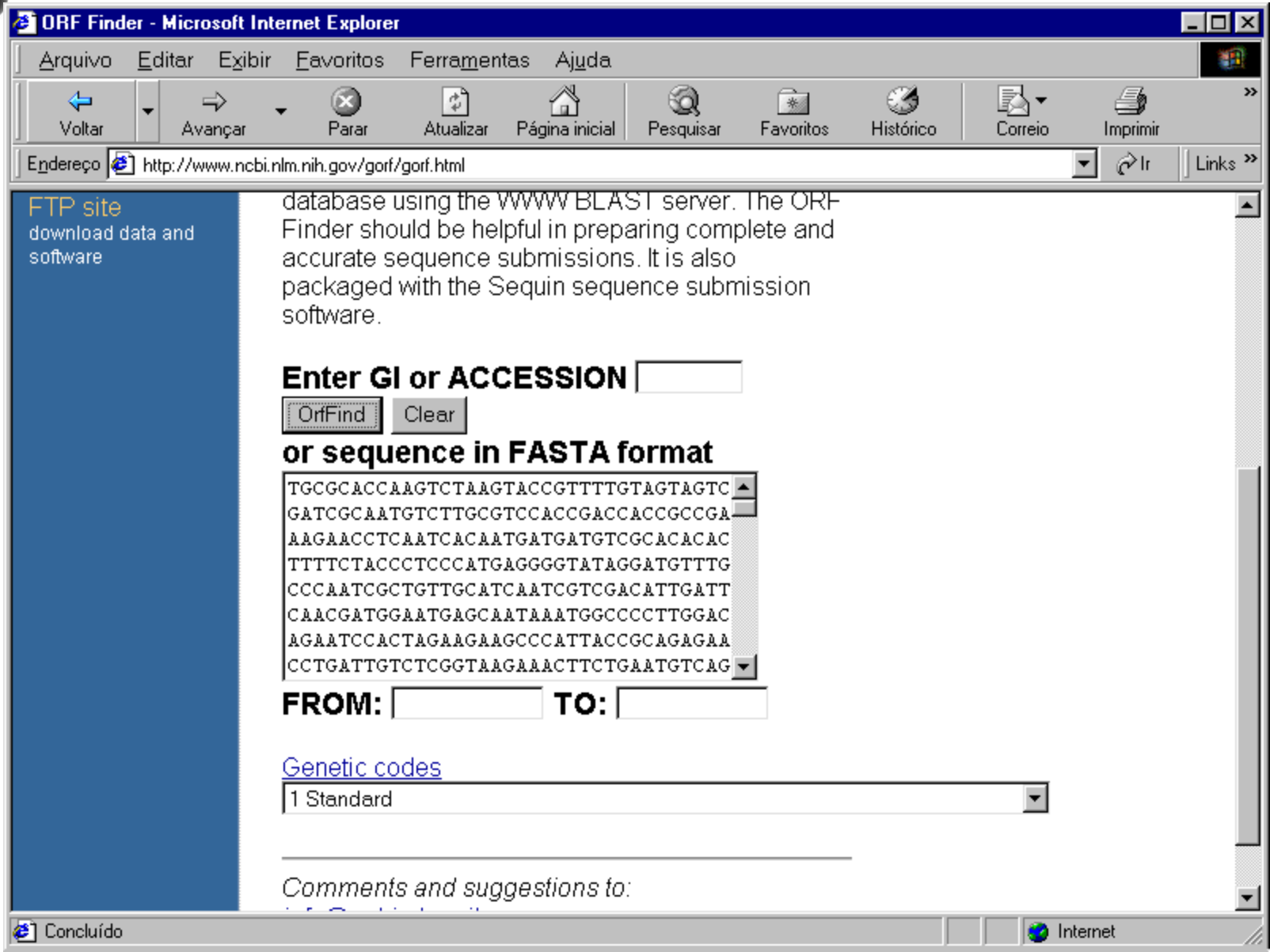
```
Query: 1   NFEGGQAVYELDRGGHHDHMVDVDTGHVIEFESEEIEALQRQIAAKHGYELEEHSLVLYV 60
           NF+GG AV+EL  GGHHDHMV+V+T  VIEF   EIE  QR+I A+HG+EL +H+LVLYV
Sbjct: 71  NFDGGHAVFELADGGHHDHMVNVETSEVIEFMDAEIEKRQREIVAEHGFELVDHNLVLYV 130

Query: 61  RKKR 64
           RKK+
Sbjct: 131 RKKK 134
```

>gi|417010|sp|Q03456|FUR_PSEAE FERRIC UPTAKE REGULATION PROTEIN (FERRIC UPTAKE REGULATOR)
Length = 134





ORF Finder

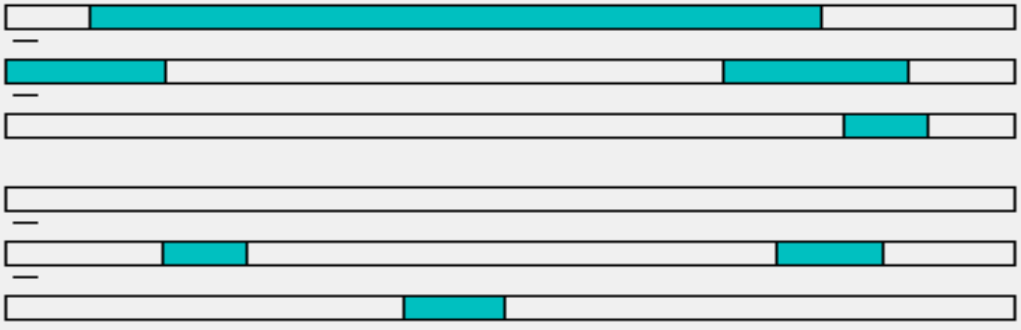

ORF Finder - Microsoft Internet Explorer
Arquivo Editar Exibir Favoritos Ferramentas Ajuda
Voltar Avançar Parar Atualizar Página inicial Pesquisar Favoritos Histórico Correio Imprimir
Endereço http://www.ncbi.nlm.nih.gov/gorf/gorf.html
Ir Links
FTP site
download data and software
database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.
Enter GI or ACCESSION
OrfFind
Clear
or sequence in FASTA format
TGCGCACCAAGTCTAAGTACCGTTTTGTAGTAGTC
GATCGCAATGTCTTGCGTCCACCGACCACCGCCGA
AAGAACCTCAATCACAATGATGATGTCGCACACAC
TTTTCTACCCTCCCATGAGGGGTATAGGATGTTTG
CCCAATCGCTGTTGCATCAATCGTCGACATTGATT
CAACGATGGAATGAGCAATAAATGGCCCCTTGGAC
AGAATCCACTAGAAGAAGCCCATTACCGCAGAGAA
CCTGATTGTCTCGGTAAGAAACTTCTGAATGTCAG
FROM: TO:
Genetic codes
1 Standard
Comments and suggestions to:
Concluído
Internet

ORF Finder (Open Reading Frame Finder)

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

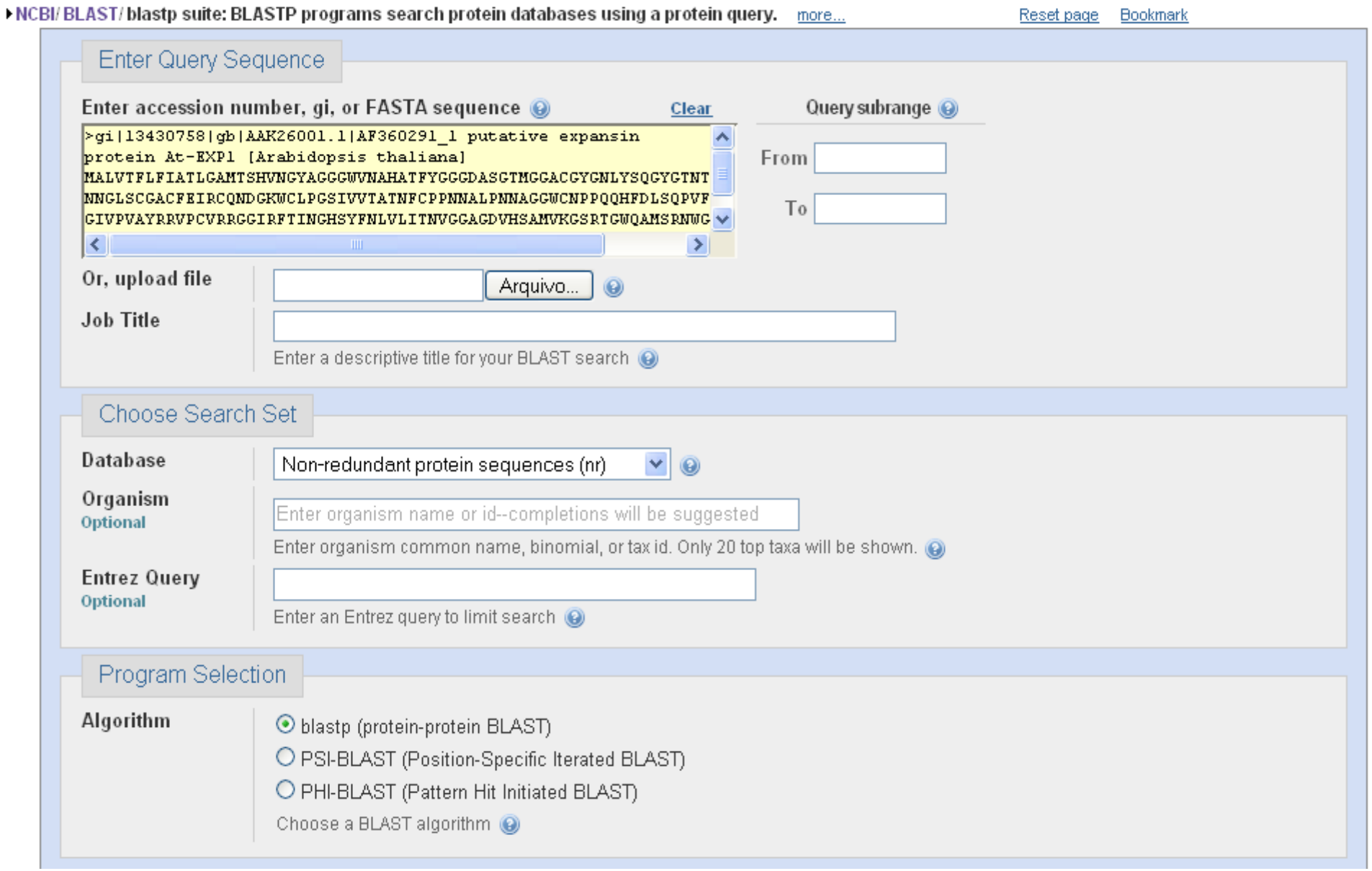
Anonymous

View | 1 GenBank | Redraw | 100 | SixFrames



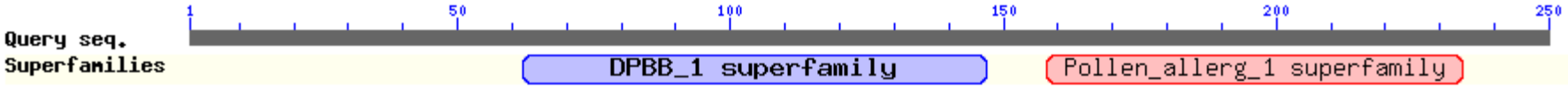
Frame	from	to	Length
+1	121	1137	1017
+2	1001	1258	258
+2	2	223	222
-2	1075	1221	147
-3	555	698	144
+3	1167	1286	120
-2	220	336	117

Blast P


NCBI/ BLAST/ blastp suite: BLASTP programs search protein databases using a protein query.
more...
Reset page
Bookmark
Enter Query Sequence
Enter accession number, gi, or FASTA sequence
Clear
Query subrange
>gi|13430758|gb|AAK26001.1|AF360291_1 putative expansin
protein At-EXP1 [Arabidopsis thaliana]
MALVTFLFIATLGAMTSHVNGYAGGGWVNAHATFYGGGDASGTMGGACGYGNLYSQGYGTNT
NNGLSCGACFEIRCQNDGKWCLPGSIVVTATNFCPPNNALPNNAGGWCNPPQQHFDLSQPVF
GIVPVAYRRVPCVRRGGIRFTINGHSYFNLVLITNVGGAGDVHSAMVKGSRTGWQAMSRNWG
From
To
Or, upload file
Arquivo...
Job Title
Enter a descriptive title for your BLAST search
Choose Search Set
Database
Non-redundant protein sequences (nr)
Organism
Optional
Enter organism name or id--completions will be suggested
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Entrez Query
Optional
Enter an Entrez query to limit search
Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
Choose a BLAST algorithm

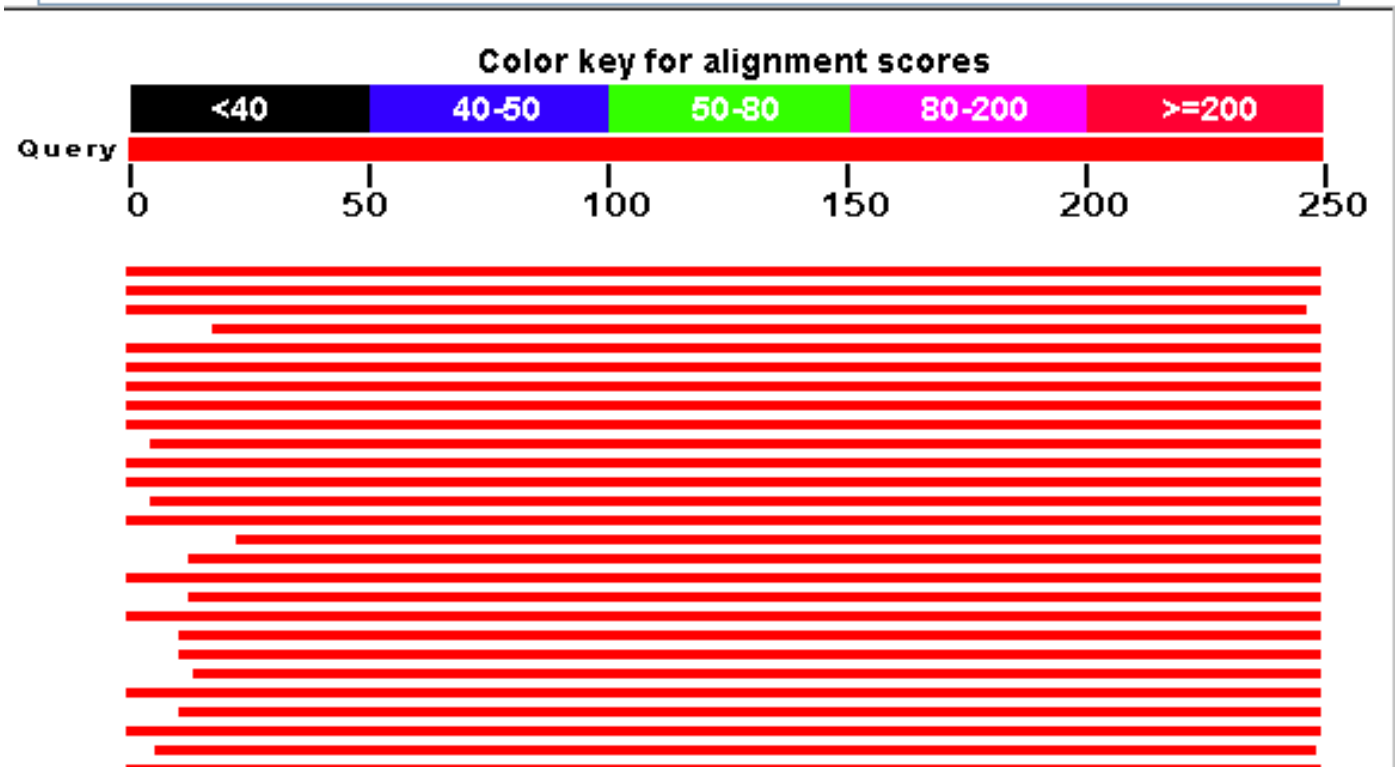
Blast P

Putative conserved domains have been detected, click on the image below for detailed results.



Request ID	7DM5VSVE012
Status	Searching
Submitted at	Thu Jul 10 23:17:15 2008
Current time	Thu Jul 10 23:17:22 2008
Time since submission	00:00:07

Mouse-over to show defline and scores, click to show alignments



GENE ID: 843288 ATEXPA1 | ATEXPA1 (ARABIDOPSIS THALIANA EXPANSIN A1)
[Arabidopsis thaliana] **(10 or fewer PubMed links)**

Score = 503 bits (1296), Expect = 3e-141, Method: Compositional matrix adjust.
Identities = 246/250 (98%), Positives = 247/250 (98%), Gaps = 0/250 (0%)

```
Query  1    MALVTFLFIATLGAMTSHVNGYAGGGWVNAHATFYGGGDASGTMGGACGYGNLYSQGYGT  60
            MALVTFLFIATLGAMTSHVNGYAGGGWVNAHATFYGGGDASGTMGGACGYGNLYSQGYGT
Sbjct  1    MALVTFLFIATLGAMTSHVNGYAGGGWVNAHATFYGGGDASGTMGGACGYGNLYSQGYGT  60

Query  61   NTAALSTALFNNGLSCGACFEIRCQNDGKWCLPGSIVVTATNFCPPNNALPNNAGGWCNP  120
            NTAALSTALFNNGLSCGACFEIRCQNDGKWCLPGSIVVTATNFCPPNNALPNNAGGWCNP
Sbjct  61   NTAALSTALFNNGLSCGACFEIRCQNDGKWCLPGSIVVTATNFCPPNNALPNNAGGWCNP  120

Query  121  PQQHFDLSQPVFQRIAQYRAGIVPVAYRRVPCVRRGGIRFTINGHSYFNLVLITNVGGAG  180
            PQQHFDLSQPVFQRIAQYRAGIVPVAYRRVPCVRRGGIRFTINGHSYFNLVLITNVGGAG
Sbjct  121  PQQHFDLSQPVFQRIAQYRAGIVPVAYRRVPCVRRGGIRFTINGHSYFNLVLITNVGGAG  180

Query  181  DVHSAMVKGSRTGWQAMSRNWGQNWQSNSYLNGQSLSFKVTTSDGQTIVSNNVANAGWSF  240
            DVHSAMVKGSRTGWQAMSRNWGQNWQSNSYLNGQSLSFKVTTSDGQTIVSNNVANAGWSF
Sbjct  181  DVHSAMVKGSRTGWQAMSRNWGQNWQSNSYLNGQSLSFKVTTSDGQTIVSNNVANAGWSF  240

Query  241  GQTFTGAQLR  250
            GQTFT   +R
Sbjct  241  GQTFTVEAVR  250
```

>gb|AAU90318.1| Alpha-expansin 1 precursor , putative [Solanum demissum]
Length=249

Score = 400 bits (1028), Expect = 4e-110, Method: Compositional matrix adjust.
Identities = 203/250 (81%), Positives = 219/250 (87%), Gaps = 1/250 (0%)

```
Query  1    MALVTFLFIATLGAMTSHVNGYAGGGWVNAHATFYGGGDASGTMGGACGYGNLYSQGYGT  60
            MA     F+  L AM S V GY GGGW+NAHATFYGGGDASGTMGGACGYGNLYSQGYGT
Sbjct  1    MAYFGICFVGLL-AMVSSVYGYGGGGWINAHATFYGGGDASGTMGGACGYGNLYSQGYGT  59

Query  61   NTAALSTALFNNGLSCGACFEIRCQNDGKWCLPGSIVVTATNFCPPNNALPNNAGGWCNP  120
            NTAALSTA+FNNGLSCG+CFE+RC ND + CLPGSIVVTATNFCPPNNALPNNAGGWCNP
Sbjct  60   NTAALSTAMFNNGLSCGSCFELRCVNDRQGCLPGSIVVTATNFCPPNNALPNNAGGWCNP  119

Query  121  PQQHFDLSQPVFQRIAQYRAGIVPVAYRRVPCVRRGGIRFTINGHSYFNLVLITNVGGAG  180
            P  HFDLSQP+FQ IA Y+AGIVPVAYRRVPC RRGGIRFTINGHSYFNLVL+TNVGG G
Sbjct  120  PLHHFDLSQPIFQHIAHYKAGIVPVAYRRVPCRRRGGIRFTINGHSYFNLVLVTNVGGGG  179

Query  181  DVHSAMVKGSRTGWQAMSRNWGQNWQSNSYLNGQSLSFKVTTSDGQTIVSNNVANAGWSF  240
            DVHS  VKGSRTGWQ MSRNWGQNWQSN+ LNGQ+LSFKVTT DG++++S NVA A WSF
Sbjct  180  DVHSVAVKGSRTGWQPMSRNWGQNWQSNNNLNGQTLSFKVTTGDGRSLISYNVAPAHWSF  239

Query  241  GQTFTGAQLR  250
            GQT+TGAQ
Sbjct  240  GQTYTGAQFH  249
```

Multiple Sequence Alignment by CLUSTALW

CLUSTALW | MAFFT | PRRN

Help

General Setting Parameters:

Output Format: CLUSTAL

Pairwise Alignment: **FAST/APPROXIMATE** **SLOW/ACCURATE**

Enter your sequences (with labels) below (copy & paste): **PROTEIN** **DNA**

Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

```
> Arabidopsis thaliana
MRKSFKDSLKALEADIQFANTLASEYPEEYDGGYVQMRLSYSPAAHLFLFLLQWTDCHF
AGALGLLRILI
YKAYVDGKTTMSLHERKTSIREFYDVLFPSLLQLHGGITDVEERKQKEICDKRYRKKDR
TDKGKMSEIDL
```

Or give the file name containing your query

Procurar...

Execute Multiple Alignment | Reset

CLUSTALW Result

[clustalw.aln][clustalw.dnd]

```
 CLUSTAL W (1.83) Multiple Sequence Alignments


Sequence type explicitly set to Protein
Sequence format is Pearson
Sequence 1: _Arabidopsis       242 aa
Sequence 2: _Oryza             172 aa
Sequence 3: _Arabidopsis2      172 aa
Sequence 4: _Medicago          243 aa
Sequence 5: _Avicennia         207 aa
Sequence 6: _Xenopus           242 aa
Start of Pairwise alignments
Aligning...


Sequences (1:2) Aligned. Score: 56.9767
Sequences (1:3) Aligned. Score: 100
Sequences (1:4) Aligned. Score: 61.9835
Sequences (1:5) Aligned. Score: 61.8357
Sequences (1:6) Aligned. Score: 14.4628
Sequences (2:2) Aligned. Score: 100
Sequences (2:3) Aligned. Score: 37.7907
Sequences (2:4) Aligned. Score: 54.0698
Sequences (2:5) Aligned. Score: 38.9535
Sequences (2:6) Aligned. Score: 13.9535
Sequences (3:2) Aligned. Score: 37.7907
Sequences (3:3) Aligned. Score: 100
```

clustalw.aln

```
CLUSTAL W (1.83) multiple sequence alignment


_Arabidopsis       --------MRKSFKDSLKALEADIQFANTLASEYPEEYDGGYVQMRLSYSPAAH--LFLF
_Arabidopsis2      --------MRKSFKDSLKALEADIQFANTLASEYPEEYDGGYVQMRLSYSPAAH--LFLF
_Avicennia         ----MLRVMKKSFKDSLKALEADIQHANTLASDYPTEHDGACLQMRLSYSPCAH--LFLF
_Medicago          -------MGKSLFQESLKALEADIQYANTLALGHPRDKEGGCFQMRLSYSPVAP--LFLS
_Oryza             ------------------------------------------------------------
_Xenopus           MGQQISSQTQTVINKLPEKLVKHMTLVRESGVLTYEEFLGRVAELNDLTAKLASGQEKHL


_Arabidopsis       LLQWTDCHFAGALGLLRILIYKAYVDGKTTMSLHERKTSIREFYDVLFPSLLQLHGGITD
_Arabidopsis2      LLQWTDCHFAGALGLLRILIYKAYVDGKTTMSLHERKTSIREFYDVLFPSLLQLHGGITD
_Avicennia         LVQWADCHLAGVLGLIRILIYKAYEDGKTTRSICERKASLREFYGVIFPSLLQLHRGITD
_Medicago          LVQWTDYRLAGALGLLRILIYVTYGNGKTTISIYERKASIRQFYSIIFPALLQLQKGVTD
_Oryza             --------------------YKVYVDGTTTMSTHERKASIKEFYAVIFPSLLQLQRGITD
_Xenopus           LFEVQPGSDSSALWKVAVRVLCTKINKTTGIVEISRIMNLYQFMQLYKDITSHAAGVFTQ
                                       .  : .*     .*  .: :*  :      :    .*:

_Arabidopsis       VEER--------KQKEICDKR-YRKKDRTDKGKMSEIDLEREEECGICLEIRNKVVLPTC
_Arabidopsis2      VEER--------KQKEICDKR-YRKKDRTDKGKMSEIDLEREEECGICLEIRNKVVLPTC
_Avicennia         VEER--------KQRVIIPQLNTRRRDEMAKGKLSEIEIEREEECAICMEMNSKVVLPSC
_Medicago          LEER--------KQKEVYANR-YQKKTDFKDRRESKIDIEREKECGVCLEVKAKVVLPNC
_Oryza             TEDK--------KQKAVCMER-YRRRDEDERNILSEIDAEREEECGICMEMNSKVVLPNC
_Xenopus           SSSGGGGEGGGGAAEENTESFNSLSSCQASFLMGKVKQLTDEEECCICMDGRADLILP-C
                    ..           .    .              .  :   *:** :*:: . .::** *

_Arabidopsis       NHSMCINCYRNWRARSQSCPFCRGSLKRVNSGDLWIYTCSAEIADLPAIYKENLKRLLIY
_Arabidopsis2      NHSMCINCYRN-------------------------------------------------
_Avicennia         SHSMCMKCYRNWRARFSVVPVLSRQSKEDEFWGALDLHKPL-------------------
_Medicago          CHQMCFKCYREWCLRSQSCPFCRDSLKRVNSGDLWIYTDTSDIVDVGTIFKENCKILFLY
_Oryza             THNMCLRCYQDWNSRSQSCPFCRDNLKKTDPGDLWIYVEDQDVVDLETVSRENLRRLFMY
_Xenopus           AHSFCQKCIDKWSDRNRNCPICRLQVTGAN--DSWVVSDAPTDEDVAS-----------Y
                    *.:* .*  .
```

Exemplo: Alinhamento gerado pelo ClustalW

```
CLUSTAL W (1.83) multiple sequence alignment

gi|662381|gb|AAC41413.1|      MENKVRFKLHKVKKNWVTIGVTTLSMVALAGGSLLAQGKVEADETSAPNG 50
gi|662379|gb|AAC41412.1|      MDKKVHYKMHKVKKQWVTIAVTGLSLGAVSAVSLGTNDGVVQADEHTDAT 50
                              *::**::*:*****:*****.** **: *::. ** ::: *   :  :

gi|662381|gb|AAC41413.1|      DGLQQLSEDGTASLVTTTTVTEQASAQASVSAVATASVSHETSFQAATSA 100
gi|662379|gb|AAC41412.1|      VAIPDITVDTGTVSNDTTAAQDPTTAVAATNDVATDQATPTATFDLTTDT 100
                               .: ::: *  :    **:. : ::* *:.. *** ..:  ::*: :*.:

gi|662381|gb|AAC41413.1|      VSQEATAQAQTSPVASQEVAVSSQTQSSGQETQTTEQVSQGQTSTQVAGQ 150
gi|662379|gb|AAC41412.1|      TNTVAANAVDTVATVGTDRAATTNDTTATNDTAVDTTNNNTTTDTTTVTD 150
                              ..  *:  .:* .... : *.:::  :: ::* .    .:  *.* .. :

gi|662381|gb|AAC41413.1|      TSAQSTPSVTEQARPRVLTNAAPAIATRAADSTIRINANRNTNITITASG 200
gi|662379|gb|AAC41412.1|      RAATTERRATGARRGPTGGRRATPVNGNTNNANNTVTVVNNDLPATNNVV 200
                               :* :   .*   *  .  . *..:  .: ::.  :.. .*   : .

gi|662381|gb|AAC41413.1|      TTPNVTIITGPNTPKPNVTVTSPNGTRPNVTIVTQPNQPNKPVQPSQPSQ 250
gi|662379|gb|AAC41412.1|      TDG------------------------------------------------ 203
                              *

gi|662381|gb|AAC41413.1|      PNKPVQPNQPSLDYKPVASNLKTIDGKQYYVENGVVKKNAAIELDGRLYY 300
gi|662379|gb|AAC41412.1|      ------PSHIKTINGKQYYVEDDGTIRKNYVLERIGGSQYFNAETGELSN 247
                                    *.: :          :    :: ** : :  ::      *.*

gi|662381|gb|AAC41413.1|      FDETGAMVDQSKPLYRADAIPNNSIYAVYNQAYDTSSKSFEHLDNFLTAD 350
gi|662379|gb|AAC41412.1|      QKEYRFDKNGGTGSSADSTNTNVTVNGDKNAFYGTTDKDIELVDGYFTAN 297
                               .*     : ..     .: .* :: .  *  *.*:.*.:* :*.::**:

gi|662381|gb|AAC41413.1|      SWYRPKQILKDGKNWTASTEKDYRPLLMTWWPDKVTQVNYLNYMSQQGFG 400
gi|662379|gb|AAC41412.1|      TWYRPKEILKDGKEWTASTENDKRPLLTVWWPSKAIQASYLNYMKEQGLG 347
                              :******:******:******:* **** .***.*. *..******.:**:*
```

Alinhamento de partes de duas proteínas Glucosyltransferase gerado pelo ClustalW

- **Combinações idênticas → (*)**
- **Similaridades → (":" e ".")**
- ***Gaps* → (-) para cada posição**

```
_Arabidopsis2      NHSMCINCYRN--------------------------------------------------
_Avicennia         SHSMCMKCYRNWRARFSVVPVLSRQSKEDEFWGALDLHKPL--------------------
_Medicago          CHQMCFKCYREWCLRSQSCPFCRDSLKRVNSGDLWIYTDTSDIVDVGTIFKENCKILFLY
_Oryza             THNMCLRCYQDWNSRSQSCPFCRDNLKKTDPGDLWIYVEDQDVVDLETVSRENLRRLFMY
_Xenopus           AHSFCQKCIDKWSDRNRNCPICRLQVTGAN--DSWVVSDAPTDEDVAS-----------Y
                    *.:* .*  .

_Arabidopsis       IDKLPLVTSDPNLVPYAPLPR
_Arabidopsis2      ---------------------
_Avicennia         ---------------------
_Medicago          IEKLPLIIPDPRHVSYDPFFR
_Oryza             INKLPLIVPDVIFSIYDSHIK
_Xenopus           ILNLADEVGQPNDTFF-----
```

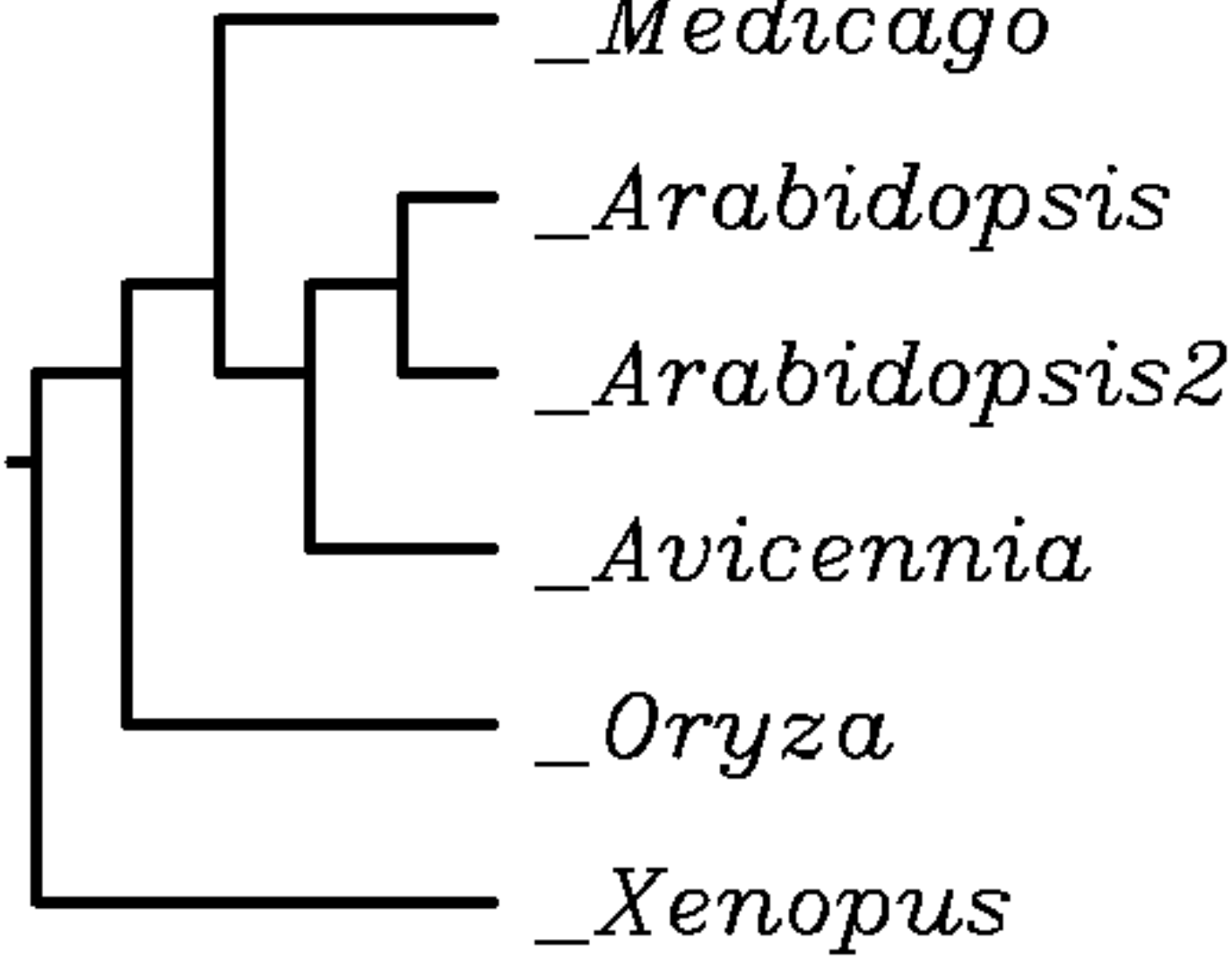
---

clustalw.dnd

```
(
(
(
_Arabidopsis:-0.01079,
_Arabidopsis2:0.01079)
:0.13169,
_Avicennia:0.20448)
:0.07866,
(
_Oryza:0.25899,
_Xenopus:0.60148)
:0.03373,
_Medicago:0.19544);
```

Select tree menu Exec

Select tree menu
N-J Tree
N-J Tree with branch length
Unrooted N-J Tree
Dendrogram
Dendrogram with branch length


_Medicago
_Arabidopsis
_Arabidopsis2
_Avicennia
_Oryza
_Xenopus

TRAPID: Rapid Analysis of Transcriptome Data

TRAPID

TRAPID is an online tool for the fast, reliable and user-friendly analysis of de novo transcriptomes

Through a highly optimized processing pipeline the TRAPID system offers functional and comparative analyses for transcriptome data sets. TRAPID is highly competitive with respect to other existing solutions with regards to both speed and quality.

TRAPID features

- Allows each user to have up to 10 different working sets, each allowing up to a 200,000 putative transcripts
- Allows the user to select a reference database of choice; currently >170 genomes are available through PLAZA 2.5 and OrthoMCLDB version 5.0
- Assign each transcript to a reference gene family or orthologous group.
- Transfer functional annotation based on homology/orthology information for each transcript
- Perform gene family-based analyses such as multiple sequence alignments and phylogenetic tree construction
- Performs functional GO enrichment analysis of subsets
- Extensive editing and export capabilities
- Free of charge for academic use

Plaza

Quick Selection

Available PLAZA versions *(select the most appropriate)*

PLAZA 3.0 Dicots (2014)

Summary

Latest iteration of the PLAZA platform. Focused on dicot species, but with some others as reference organisms.

This platform may still contain minor bugs. Do not hesitate to contact us if you find any!

PLAZA 3.0: an access point for plant comparative genomics

Go to PLAZA 3.0 Dicots

Species Included ›

PLAZA 3.0 Monocots (2014)

Summary

Latest iteration of the PLAZA platform. Focused on monocot species, but with some others as reference organisms.

This platform may still contain minor bugs. Do not hesitate to contact us if you find any!

PLAZA 3.0: an access point for plant comparative genomics

Go to PLAZA 3.0 Monocots

Species Included ›

The Arabidopsis Information Resource

The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model higher plant *Arabidopsis thaliana* . Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community. Gene product function data is updated every two weeks from the latest published research literature and community data submissions. Gene structures are updated 1-2 times per year using computational and manual methods as well as community submissions of new and updated genes. TAIR also provides extensive linkouts from our data pages to other Arabidopsis resources.

The Arabidopsis Biological Resource Center at The Ohio State University collects, reproduces, preserves and distributes seed and DNA resources of *Arabidopsis thaliana* and related species. Stock information and ordering for the ABRC are fully integrated into TAIR.

CARNEGIE INSTITUTION FOR SCIENCE

TAIR is located at the Carnegie Institution for Science Department of Plant Biology and funded by the National Science Foundation.

Updates on TAIR funding are available **here**.

Breaking News

Subscribe to news feed

Follow our Twitter feed

Join our Facebook group

TAIR10 Genome Release [Nov 17, 2010]
The TAIR10 genome release is now available, with 126 new loci, updates to 1184 gene structures and 2099 new gene models

ABRC Ordering Policy [July 27, 2010]
Effective August 1, 2010 we can no longer process orders that have both seed and DNA stocks. Please make separate orders.

ASPB Early Bird Registration Deadline [April 30, 2010]
The Early Bird Registration Deadline of this year's Plant Biology Meeting in Montreal is

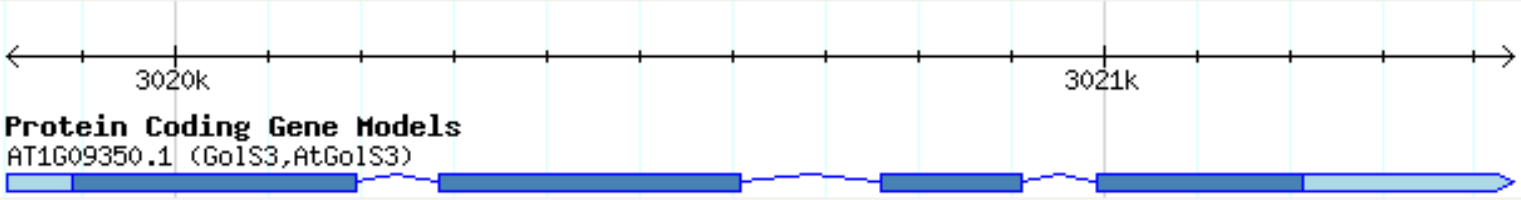
Check All | Uncheck All

	Locus	Description	Gene Model(s)	Other Names	Keywords
1	AT1G09350	galactinol synthase 3 (GolS3); CONTAINS InterPro DOMAIN/s: Glycosyl transferase, family 8 (InterPro:IPR002495); BEST Arabidop....	AT1G09350.1	ARABIDOPSIS THALIANA GALACTINOL SYNTHASE 3 ATGOLS3 F14J9.1 F14J9_1 GALACTINOL SYNTHASE 3 GOLS3	carbohydrate biosynthetic process, cellular component unknown, response to cold, transferase activity, transferase activity, transferring glycosyl groups, transferase activity, transferring hexosyl groups, transferring glycosyl groups, transferring hexosyl groups
2	AT1G56600	galactinol synthase 2 (GolS2); CONTAINS InterPro DOMAIN/s: Glycosyl transferase, family 8 (InterPro:IPR002495); BEST Arabidop....	AT1G56600.1	ARABIDOPSIS THALIANA GALACTINOL SYNTHASE 2 ATGOLS2 F25P12.95 F25P12_95 GALACTINOL SYNTHASE 2 GOLS2	carbohydrate biosynthetic process, cellular component unknown, response to water deprivation, transferase activity, transferase activity, transferring glycosyl groups, transferase activity, transferring hexosyl groups, transferring glycosyl groups, transferring hexosyl groups
3	AT1G60450	galactinol synthase 7 (GolS7); CONTAINS InterPro DOMAIN/s: Glycosyl transferase, family 8 (InterPro:IPR002495); BEST Arabidop....	AT1G60450.1	ARABIDOPSIS THALIANA GALACTINOL SYNTHASE 7 ATGOLS7 GALACTINOL	carbohydrate biosynthetic process, cellular component unknown, transferase activity, transferase activity, transferring glycosyl groups, transferase activity, transferring hexosyl groups, transferring glycosyl groups, transferring hexosyl groups

Gene Model: AT1G09350.1 [Help]

Date last modified	2010-11-17
Name	AT1G09350.1
Name Type	orf
Gene Model Type	protein_coding
TAIR Accession	Gene:2012319
Description	galactinol synthase 3 (GolS3); CONTAINS InterPro DOMAIN/s: Glycosyl transferase, family 8 (InterPro:IPR002495); BEST Arabidopsis thaliana protein match is: galactinol synthase 2 (TAIR:AT1G56600.1); Has 1260 Blast hits to 1259 proteins in 288 species: Archae - 0; Bacteria - 98; Metazoa - 258; Fungi - 282; Plants - 479; Viruses - 73; Other Eukaryotes - 70 (source: NCBI BLink).
Chromosome	1
Locus	AT1G09350 (Note: use this locus link to see all functional annotations, associated gene models, markers and ESTs).

Map Detail Image



Gene Alias

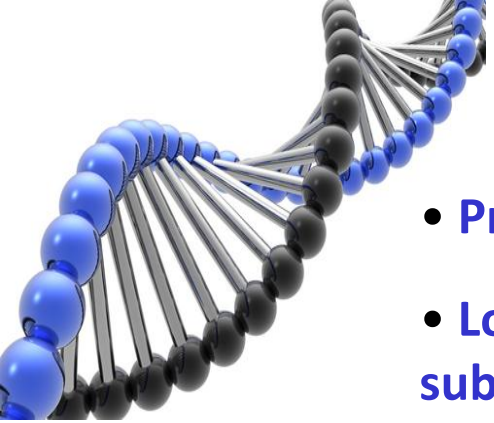
name	type
F14J9_1	orf
F14J9.1	orf

Symbols

symbol	full name
AtGolS3	ARABIDOPSIS THALIANA GALACTINOL SYNTHASE 3
GolS3	GALACTINOL SYNTHASE 3

Protein Data

name	Length	molecular	isoelectric	domains(# of domains)



Outras Ferramentas da Bioinformática

- **Predição de proteínas:** Orf Finder (NCBI)
- **Localização de peptídeo sinal e predição de localização subcelular:**

Signal P, PSORT Prediction, Target P ...

- **Indetificação de sítios de glicosilação:** SignalP ...
- **Identificação de motivos protéicos:** ScanProsite ...
- **Alinhamento:** Mega, ClustalW, Phylip, WebLogo ...
- **Formação de Contigs (Clusterização):** Mega, BioEdit, NTI Vector, CLC ...
- **Árvores Filogenéticas:** Mega, ClustalW ...

CBS – Center for Biological Sequence Analysis (http://www.cbs.dtu.dk/services/)

NUCLEOTIDE SEQUENCES

Whole genome visualization and analysis

« GenomeAtlas
DNA structural atlases for complete microbial Genomes

Gene finding and splice sites

« EasyGene
Genes in prokaryotes

HMMgene
Genes in eukaryotes

NetAspGene
Intron splice sites in Aspergillus DNA

« NetGene2
Intron splice sites in human, C. elegans and A. thaliana DNA

NetPlantGene
Intron splice sites in Arabidopsis thaliana DNA

« NetStart
Translation start in vertebrate and A. thaliana DNA

NetUTR
Splice sites in 5' UTR regions of human genes

« Promoter
Transcription start sites in vertebrate DNA

« RNAmmer
Ribosomal RNA sub units

Analysis of DNA microarray data

GenePublisher
Analysis of DNA microarray data

« OligoWiz
Design of oligonucleotides for DNA microarrays

AMINO ACID SEQUENCES

Protein sorting

ChloroP »
Chloroplast transit peptides and their cleavage sites in plant proteins

LipoP »
Signal peptidase I & II cleavage sites in gram- bacteria

NetNES
Leucine-rich nuclear export signals (NES) in eukaryotic proteins

SecretomeP »
Non-classical and leaderless secretion of proteins

SignalP »
Signal peptide and cleavage sites in gram+, gram- and eukaryotic amino acid sequences

TargetP »
Subcellular location of proteins: mitochondrial, chloroplastic, secretory pathway, or other

TatP
Twin-arginine signal peptides

Post-translational modifications of proteins

DictyOGlyc
O-(alpha)-GlcNAc glycosylation sites (trained on *Dictyostelium discoideum* proteins)

NetAcet
N-terminal acetylation in eukaryotic proteins

NetCGlyc
C-mannosylation sites in mammalian proteins

NetCorona
Coronavirus 3C-like proteinase cleavage sites in proteins

NetGlycate »
Glycation of ε amino groups of lysines in mammalian proteins

CBS – Center for Biological Sequence Analysis (http://www.cbs.dtu.dk/services/)

Post-translational modifications of proteins

DictyOGlyc
O-(alpha)-GlcNAc glycosylation sites (trained on *Dictyostelium discoideum* proteins)

NetAcet
N-terminal acetylation in eukaryotic proteins

NetCGlyc
C-mannosylation sites in mammalian proteins

NetCorona
Coronavirus 3C-like proteinase cleavage sites in proteins

NetGlycate »
Glycation of ε amino groups of lysines in mammalian proteins

NetNGlyc »
N-linked glycosylation sites in human proteins

NetOGlyc »
O-GalNAc (mucin type) glycosylation sites in mammalian proteins

NetPhos »
Generic phosphorylation sites in eukaryotic proteins

NetPhosK
Kinase specific phosphorylation sites in eukaryotic proteins

NetPhosYeast
Serine and threonine phosphorylation sites in yeast proteins

NetPicoRNA
Posttranslational cleavage by picornaviral proteases

ProP »
Arginine and lysine propeptide cleavage sites in eukaryotic protein sequences

YinOYang »
O-(beta)-GlcNAc glycosylation and *Yin-Yang* sites (intracellular/nuclear proteins)

Immunological features

BepiPred »
Linear B-cell epitopes

DiscoTope »
Discontinuous B-cell epitopes

NetChop »
Proteasomal cleavages (MHC ligands)

NetCTL »
Integrated class I antigen presentation

NetMHC »
Binding of peptides to MHC class I alleles

NetMHCII »
Binding of peptides to MHC class II alleles

NetMHCIIpan
Pan-specific binding of peptides to MHC class II HLA-DR alleles of known sequence

NetMHCpan
Pan-specific binding of peptides to MHC class I alleles of known sequence

VDJsolver »
Analysis of human immunoglobulin VDJ recombination

Protein function and structure

ArchaeaFun
Enzyme/non-enzyme and enzyme class (Archaea)

CPHmodels
Protein structure from sequence: distance constraints

distanceP
Protein distance constraints

EPipe »
Functional differences of protein variants

ProtFun
Protein functional category and enzyme class (Eukarya)

RedHom
Reduction of sequence similarity in a data set

METACYc – Enclycopedia of Metabolic Pathway - http://metacyc.org/

Quick Search

Go

Search
Database Search
Advanced Database Search

Browse
Pathways
Compounds
Reactions

Home

About MetaCyc
MetaCyc User's Guide
Help
Database Overview
Graphical Overview
Querying Examples
Literature Curation Process
Publications
Release Notes

Overview

MetaCyc is a database of nonredundant, experimentally elucidated metabolic pathways. MetaCyc contains more than 1,100 pathways from more than 1,500 different organisms. [more], and is curated from the scientific experimental literature. [more]

MetaCyc contains pathways involved in both primary [def] and secondary [def] metabolism, as well as associated compounds, enzymes, and genes. [more]

Motivation

The goal of MetaCyc is to catalog the universe of metabolism by storing a representative sample of each experimentally elucidated pathway. [MetaCyc mission]

MetaCyc is used in a variety of scientific applications, such as providing a reference data set for computationally predicting the metabolic pathways of organisms from their sequenced genomes, supporting metabolic engineering, helping to compare biochemical networks, and serving as an encyclopedia of metabolism. [scientific applications]

News

Recent publication:

- The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases, *Nucleic Acids Research* 36:D623-31 2008.

Expand All | Collapse All

Pathways
- **Biosynthesis**
- **Degradation/Utilization/Assimilation**
- **Detoxification**
- **Generation of precursor metabolites and energy**
- **Signal transduction pathways**
- **Superpathways**

Quick Search | Go | Query Page | Advanced Query Page | BioCyc Home | Report Errors or Provide Feedback

MetaCyc Class: Biosynthesis

Summary:

This class contains pathways that constitute a cell's complete spectrum of biosynthetic capacities, including the routes of synthesis of small molecules, macromolecules and organelles.

Parent Classes:

Pathways

Child Classes:

Amines and Polyamines Biosynthesis (27) ,
Amino acids Biosynthesis (95) ,
Aminoacyl-tRNA Charging (2) ,
Aromatic Compounds Biosynthesis (9) ,
Carbohydrates Biosynthesis (58) ,
Cell structures Biosynthesis (29) ,
Cofactors, Prosthetic Groups, Electron Carriers Biosynthesis (143) ,
Fatty Acids and Lipids Biosynthesis (60) ,
Hormones Biosynthesis (29) ,
Metabolic Regulators Biosynthesis (2) ,
Nucleosides and Nucleotides Biosynthesis (13) ,
Other Biosynthesis (8) ,
Secondary Metabolites Biosynthesis (276) ,
Siderophore Biosynthesis (6)

Pathways Pathway pages contain: Depiction of metabolic pathway, of chromosomal locations of pathway genes, and of regulation of pathway genes.

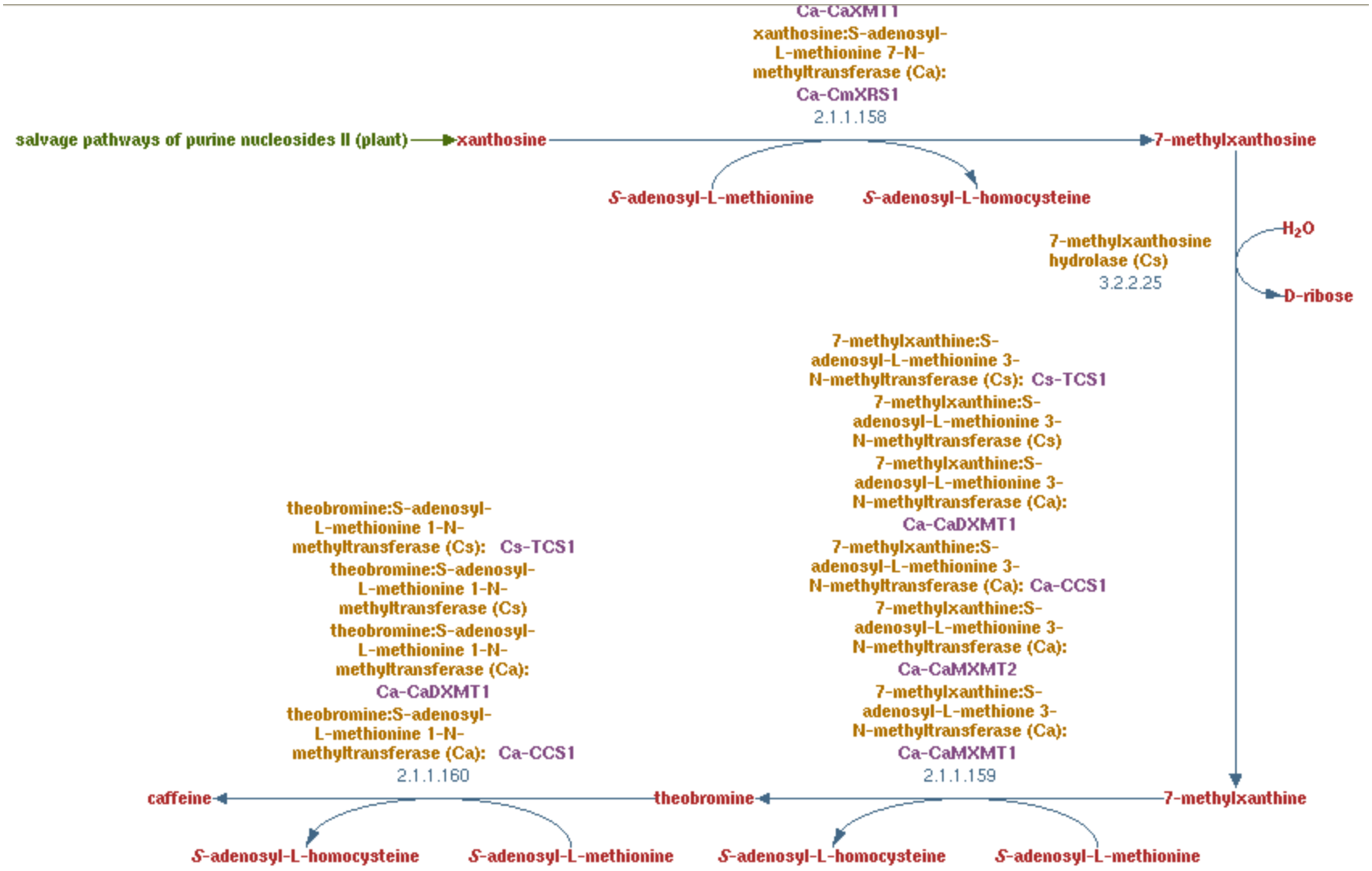
- **caffeine** biosynthesis I
- **caffeine** biosynthesis II (via paraxanthine)

Proteins Protein pages contain: Detailed comments and citations; subunit structure; cofactors, activators, and inhibitors (for enzymes), depiction of regulon (for transcription factors).

- **caffeine** synthase (polypeptide)
- **caffeine** synthase (polypeptide) - CaDXMT1
- **caffeine** synthase (polypeptide) - CCS1
- **caffeine** synthase (polypeptide) - TCS1

Compounds Compound pages contain: compound structural information, and links to all reactions and pathways in which the compound participates.

- **caffeine**


Ca-CaXMT1
xanthosine:S-adenosyl-L-methionine 7-N-methyltransferase (Ca): Ca-CmXRS1
2.1.1.158
salvage pathways of purine nucleosides II (plant)
xanthosine
7-methylxanthosine
S-adenosyl-L-methionine
S-adenosyl-L-homocysteine
H2O
7-methylxanthosine hydrolase (Cs)
3.2.2.25
D-ribose
7-methylxanthine:S-adenosyl-L-methionine 3-N-methyltransferase (Cs): Cs-TCS1
7-methylxanthine:S-adenosyl-L-methionine 3-N-methyltransferase (Cs)
7-methylxanthine:S-adenosyl-L-methionine 3-N-methyltransferase (Ca): Ca-CaDXMT1
7-methylxanthine:S-adenosyl-L-methionine 3-N-methyltransferase (Ca): Ca-CCS1
7-methylxanthine:S-adenosyl-L-methionine 3-N-methyltransferase (Ca): Ca-CaMXMT2
7-methylxanthine:S-adenosyl-L-methione 3-N-methyltransferase (Ca): Ca-CaMXMT1
2.1.1.159
theobromine:S-adenosyl-L-methionine 1-N-methyltransferase (Cs): Cs-TCS1
theobromine:S-adenosyl-L-methionine 1-N-methyltransferase (Cs)
theobromine:S-adenosyl-L-methionine 1-N-methyltransferase (Ca): Ca-CaDXMT1
theobromine:S-adenosyl-L-methionine 1-N-methyltransferase (Ca): Ca-CCS1
2.1.1.160
caffeine
theobromine
7-methylxanthine
S-adenosyl-L-homocysteine
S-adenosyl-L-methionine
S-adenosyl-L-homocysteine
S-adenosyl-L-methionine

Latest miRBase blog posts

High confidence miRNA set available for miRBase 21 By sam (July 3, 2014)

As mentioned previously, we briefly held off from releasing the set of "high confidence" miRNAs for miRBase 21, because of a last-gasp bug. Those data are now available, tagged with the label "high confidence" on the entry pages, and for download on the FTP site. The total number of miRNAs labelled "high confidence" has increased [...]

miRBase 21 finally arrives By sam (June 26, 2014)

Apologies for the longer-than-usual wait. miRBase 21 is now available on the website, and all data available for download on the FTP site. As usual, the release notes describe the major changes. Of particular note this time, the Genome Reference Consortium have released a new human genome assembly, GRCh38. We have therefore remapped the human [...]

miRBase: the microRNA database

miRBase provides the following services:

- The miRBase database is a searchable database of published miRNA sequences and annotation. Each entry in the miRBase Sequence database represents a predicted hairpin portion of a miRNA transcript (termed mir in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). Both hairpin and mature sequences are available for searching and browsing, and entries can also be retrieved by name, keyword, references and annotation. All sequence and annotation data are also available for download.
- The miRBase Registry provides miRNA gene hunters with unique names for novel miRNA genes prior to publication of results. Visit the help pages for more information about the naming service.

To receive email notification of data updates and feature changes please subscribe to the miRBase announcements mailing list. Any queries about the website or naming service should be directed at mirbase@manchester.ac.uk.

miRBase is managed by the Griffiths-Jones lab at the Faculty of Life Sciences, University of Manchester with funding from the BBSRC. miRBase was previously hosted and supported by the Wellcome Trust Sanger Institute.

miRNA count: 28645 entries

Release 21: June 2014

Search by miRNA name or keyword

Go | Example

Download published miRNA data

Download page | FTP site

Tweets

Follow

RNAcentral @RNAcentral 30 Apr

Learn how to use RNAcentral, Rfam, miRBase, and GENCODE at a 1-day course.
More on our blog:
blog.rnacentral.org/2015/04/new-tr...
pic.twitter.com/Jeg3AzBffP

Retweeted by miRBase

Tweet to @mirbase


OBRIGADO
tiagobio02@yahoo.com.br